Do causal predictors generalize better to new domains?

Vivian Y. Nastl
Max Planck Institute for Intelligent Systems, Tübingen, Germany and Tübingen AI Center
Max Planck ETH Center for Learning Systems
vivian.nastl@tuebingen.mpg.de

Moritz Hardt
Max Planck Institute for Intelligent Systems, Tübingen, Germany and Tübingen AI Center
hardt@is.mpg.de

Abstract

We study how well machine learning models trained on causal features generalize across domains. We consider 16 prediction tasks on tabular datasets covering applications in health, employment, education, social benefits, and politics. Each dataset comes with multiple domains, allowing us to test how well a model trained in one domain performs in another. For each prediction task, we select features that have a causal influence on the target of prediction. Our goal is to test the hypothesis that models trained on causal features generalize better across domains. Without exception, we find that predictors using all available features, regardless of causality, have better in-domain and out-of-domain accuracy than predictors using causal features. Moreover, even the absolute drop in accuracy from one domain to the other is no better for causal predictors than for models that use all features. In addition, we show that recent causal machine learning methods for domain generalization do not perform better in our evaluation than standard predictors trained on the set of causal features. Likewise, causal discovery algorithms either fail to run or select causal variables that perform no better than our selection. Extensive robustness checks confirm that our findings are stable under variable misclassification.

1 Introduction

The accuracy of machine learning models typically drops significantly when a model trained in one domain is evaluated in another. This empirical fact is the fruit of numerous studies [Torralba and Efros, 2011, Gulrajani and Lopez-Paz, 2020, Miller et al., 2021]. But it's less clear what to do about it. Many machine learning researchers see hope in causal modeling. Causal relationships, the story goes, reflect stable mechanisms invariant to changes in an environment. Models that utilize these invariant mechanisms should therefore generalize well to new domains [Peters et al., 2017]. The idea may be sound in theory. Intriguing theoretical results carve out assumptions under which causal machine learning methods generalize gracefully from one domain to the other [Heinze-Deml et al., 2018, Meinshausen, 2018, Schölkopf et al., 2021, Pearl and Bareinboim, 2022, Subbaswamy et al., 2022, Wang et al., 2022b].

These theoretical developments have fueled optimism about the out-of-domain generalization abilities of causal machine learning. The general sentiment is that causal methods enjoy greater external validity than kitchen-sink model fitting. In this work, we put the theorized external validity of causal machine learning to an empirical test in a wide range of concrete datasets.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

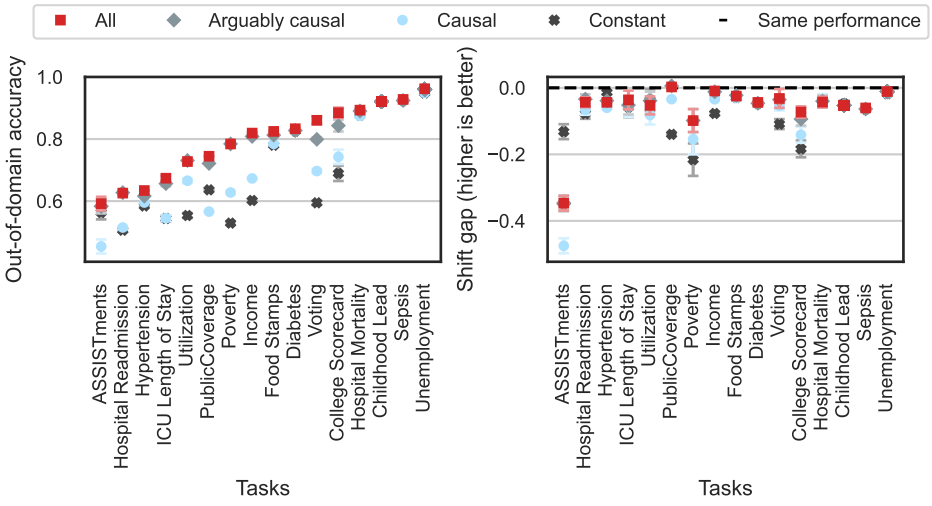


Figure 1: Best out-of-domain accuracy (left) and corresponding shift gap (right) by feature selection. Predictors based on all features have better out-of-domain accuracy than predictors using causal feature selections. Their shift gap is smaller too, up to error bars.

Our results. We consider 16 prediction tasks on tabular datasets from prior work [Ding et al., 2021, Hardt and Kim, 2023, Gardner et al., 2023] covering application settings including health, employment, education, social benefits, and politics. Each datasets comes with different domains intended for research on domain generalization. For each task we conservatively select a set of *causal features*. Causal features are those that we most strongly believe have a causal influence on the target of prediction. We also select a more inclusive set of *arguably causal* variables that include variables that may be considered causal depending on modeling choices. For each task, we compare the performance of machine learning methods trained on causal variables and arguably causal variables with those trained on all available features. In all 16 tasks, our primary finding can be summarized as:

Predictors using all available features, regardless of causality, have better in-domain and out-of-domain accuracy than predictors using causal features.

Across 16 datasets, we were unable to find a single example where causal predictors generalize better to new domains than a standard machine learning model trained on all available features. Figure 1 summarizes the situation. In greater detail, our empirical results are:

- Using all features Pareto-dominates both causal selections, with respect to in-domain and out-of-domain accuracy (up to error bars). We provide a closer look at the Pareto-frontiers of four representative tasks in Figure 2.
- The inclusive selection of arguably causal features Pareto-dominates the conservative selection of causal features, with respect to in-domain and out-of-domain accuracy (up to error bars).
- The absolute drop in accuracy from one domain to the other is smaller for all features than for causal features.
- Adding anti-causal features—i.e., features caused by the target variable—to the set of causal features improves out-of-domain performance.
- Special-purpose causal machine learning methods, such as IRM and REx, typically perform within the range of standard models trained on the conservative and inclusive selection of causal features.
- Classic causal discovery algorithms, like PC and ICP, do not provide causal parent estimates that improve upon the inclusive selection of causal features.
- Extensive robustness checks confirm that our findings are stable under misclassifications of single features.

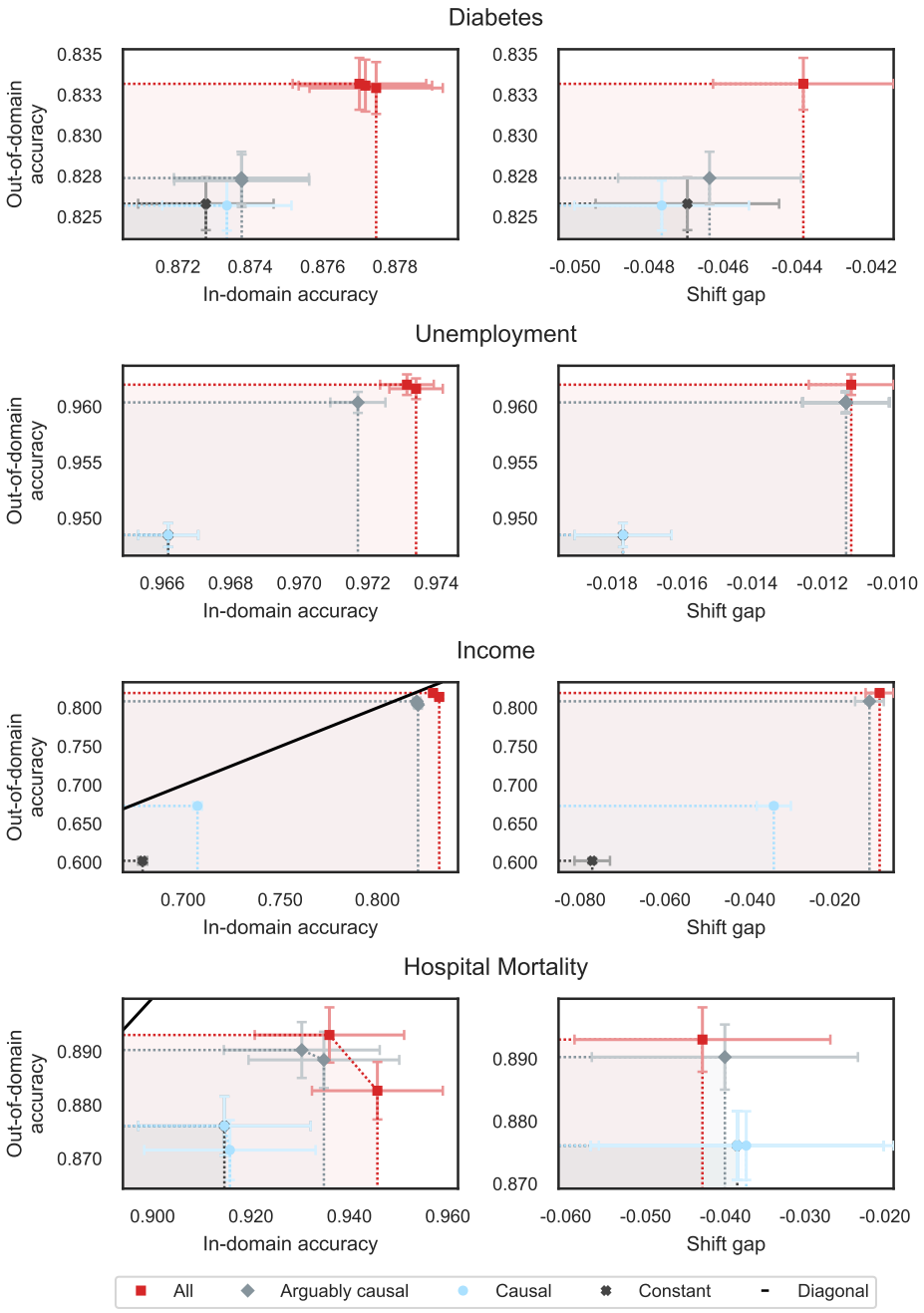


Figure 2: (Left) Pareto-frontiers of in-domain and out-of-domain accuracy by feature selection. (Right) Pareto-frontiers of shift gap and out-of-domain accuracy by feature selection. Predictors using all features Pareto-dominate predictors using causal features, with respect to in-domain and out-of-domain accuracy. Other tasks are in Appendix C.

To be sure, our findings don't contradict the theory. Rather, they point at the fact that the assumptions of existing theoretical work are unlikely to be met in the tabular data settings we study. It is, of course, always possible that those causal prediction techniques yield better results on other datasets. From this perspective, our study suggests that the burden of proof is on proponents of causal techniques to provide real benchmark datasets where these methods succeed. On the many datasets we investigated, it proved infeasible to make use of causal techniques for better out-of-domain generalization.

1.1 Related work

Existing work in causal machine learning relies on the assumption of the invariance of causal mechanisms [Haavelmo, 1944, Aldrich, 1989, Hoover, 1990, Pearl, 2009, Schölkopf et al., 2012]. The conditional distribution of the target, given the complete set of its direct causal parents, shall remain identical under interventions on variables other than the target itself. In their influential work, Peters et al. [2016] utilize this invariance property for causal discovery. In further works, it is extended to non-linear models [Heinze-Deml et al., 2018], and discovery of invariant features [Rojas-Carulla et al., 2018]. To overcome the computational burden in high-dimensional settings, Arjovsky et al. [2019] propose Invariant Risk Minimization (IRM), which learns an invariant representation of the features instead of selecting individual features. Rosenfeld et al. [2021b] however identify major failure cases of IRM. In response, multiple extensions of IRM have been proposed [Krueger et al., 2021, Wang et al., 2022a, Ahuja et al., 2022, Jiang and Veitch, 2022, Chen et al., 2023]. Another line of research assumes graphical knowledge to remove variables or apply independence constraints for regularization [Subbaswamy and Saria, 2018, Subbaswamy et al., 2019, Kaur et al., 2022, Salaudeen and Koyejo, 2024]. We refer the reader to Kaddour et al. [2022] for an overview. Aside from causal learning approaches, various domain generalization algorithms and distributional robustness methods have been developed [Ajakan et al., 2015, Sun et al., 2015, Sun and Saenko, 2016, Li et al., 2018, Levy et al., 2020, Sagawa et al., 2020, Xu et al., 2020, Zhang et al., 2021]. Each method assumes a unique type of (untestable) invariance across domains.

Gulrajani and Lopez-Paz [2020] conduct extensive experiments on image datasets to compare the performance of domain generalization algorithms, including the causal methods IRM and Risk Extrapolation (REx) [Krueger et al., 2021], in realistic settings. They find that no domain generalization methods systematically outperforms empirical risk minimization. Recently, Gardner et al. [2023] demonstrate a similar behavior for tabular data.
In our work, we shift the focus from the out-of-domain performance of specific causal machine learning *algorithms* to the performance of causal *feature sets*.

1.2 Theoretical background and motivation

To frame our empirical study, we recall some relevant theoretical background first. A *domain* $\mathcal{D}$ is composed of samples $(x_i, y_i) \sim P$, where $x_i \in \mathcal{X} \subset \mathbb{R}^p$ are the features and $y \in \mathcal{Y} \subset \mathbb{R}$ is the target [Wang et al., 2022b]. Let X and Y denote the random variables corresponding to the features and the target.

We are given m training domains $\mathcal{D}^{\text{train}} = \{\mathcal{D}^d : d = 1, \dots, m\}$. The joint distributions of features and target differ across domains, i.e. $P^d \neq P^e$ for $d \neq e$. Our goal is to learn a prediction f_θ from the training domains $\mathcal{D}^{\text{train}}$ that achieves minimum prediction error on an *unseen* test domain $\mathcal{D}^{\text{test}}$,

$$\theta^* = \arg\min_\theta \mathbb{E}_{P^{\text{test}}}[\ell(Y, f_\theta(X))], \tag{1}$$

where $\ell(\cdot, \cdot)$ is some loss function. We can compose the objective into two parts

$$\mathbb{E}_{P^{\text{train}}}[\ell(Y, f_\theta(X))] - \Delta, \tag{2}$$

where $\Delta = \mathbb{E}_{P^{\text{train}}}[\ell(Y, f_\theta(X))] - \mathbb{E}_{P^{\text{test}}}[\ell(Y, f_\theta(X))]$ is the *shift gap*. Hence, we aim to learn a classifier with the best trade-off between predicting accurately and having a low shift gap. In our empirical work, we measure the shift gap as the difference in accuracy,

$$\Delta_{\text{acc}} = \text{acc}(f_\theta, \mathcal{D}^{\text{test}}) - \text{acc}(f_\theta, \mathcal{D}^{\text{train}}). \tag{3}$$

Distributional robustness of causal mechanisms. Suppose we have a directed acyclic graph $G = (V, E)$ with nodes $V = \{1, \dots, q\}$, a random variable (Z, Y) and noise variables $\varepsilon \in \mathbb{R}^q$. A common assumption is that the target is described by the prediction f_θ via the coefficient θ^{causal}

$$Y \leftarrow f_{\theta^{\text{causal}}}(Z) + \varepsilon_q \,. \tag{4}$$

The invariance of the causal mechanism implies that these causal coefficients provide the robust estimator for the set of do-interventional distributions on the features [Meinshausen, 2018],

$$\theta^{\text{causal}} = \arg\min_\theta \sup_{Q \in \mathcal{Q}^{(\text{do})}} E_Q \left[\ell\left(Y, f_\theta\left(Z\right)\right)\right], \quad \mathcal{Q}^{(\text{do})} := \left\{ P^{(\text{do})}_{a, V\setminus\{q\}}; a \in \mathbb{R}^{q-1} \right\}. \tag{5}$$

To link to domain generalization, we need to assume that all causal parents of Y are included in the feature set X. We set $X = Z$ w.l.o.g. We also presume that the distribution of the testing domain is a do-intervention on the features, i.e., $P^{\text{test}} \in \mathcal{Q}^{(\text{do})}$. Intuitively, this postulates that the causal mechanism generating Y stays the same across domains, while features may encounter arbitrarily large interventions.
Then, the prediction error of the causal coefficients in the test domain is minimax optimal bounded,

$$\mathbb{E}_{P^{\text{test}}}[\ell(Y, f_{\theta^{\text{causal}}}(X))] \leq \min_{\theta} \sup_{Q \in \mathcal{Q}^{(\text{do})}} \mathbb{E}_{Q}\left[\ell\left(Y, f_{\theta}\left(X\right)\right)\right] . \tag{6}$$

Recent work in causal machine learning already pointed out that the minimum prediction error on test domains with mild interventions can be much smaller that the prediction error achieved by the causal coefficients [Rothenhäusler et al., 2020, Subbaswamy et al., 2022]. We conduct synthetic experiments similar to Rothenhäusler et al. [2020], further supporting the insights that that a strong shift is needed before causal features achieve best out-of-domain accuracy. The details on the setup and results of the synthetic experiments are provided in Appendix D.

Our empirical study complements these theoretical developments, as we evaluate domain generalization abilities of causal features in typical tabular datasets. We emphasize that we do not challenge the validity of causal theory like (6), but rather challenge how realistic their assumptions are.

2 Methodology

We conduct experiments on 16 classification tasks with natural domain shifts. They cover applications in multiple application areas, e.g., health, employment, education, social benefits, and politics. Most tasks are derived from the distribution shift benchmark for tabular data *TableShift* [Gardner et al., 2023]. Some others are from prior work [Hardt and Kim, 2023]. All tabular datasets contain interpretable personal information, e.g., age, education status, or individual's habits. Therefore, we can consult social and biomedical research on the causal relationships between features and target. To reflect existing epistemic uncertainty, we propose a pragmatic scheme to classify the relationship between features and target.

We term features that clearly have a causal influence on the target **causal**. We are conservative and only label features as causal when: (1) The feature has almost certainly a causal effect on the target, and (2) reverse causation from target to feature is hard to argue. We sort out any spuriously related or possibly anti-causal feature [Schölkopf et al., 2012]. However, we risk excluding relevant causal parents of the target.

For this reason, we propose the concept of **arguably causal features**. It is epistemically uncertain how these features are causally linked to the target. To be specific, we term a feature arguably causal when it suffices one of the following criteria: (1) The feature is a causal feature, or (2) the feature has a causal effect on the target and reverse causation is possible, or (3) it is plausible but not certain that the feature has a causal effect on the target. We exclude variables where it is implausible that they affect the target. Ideally, the arguably causal features cover all causal parents present in the dataset. We emphasize that both, causal features and arguably causal features, are merely approximations of the true causal parents based on current expert knowledge and restricted to available features. We further note that relationships between causal features and target might be confounded.

In some datasets and tasks we are also confronted with features that are plausibly **anti-causal**, that is: (1) The target has almost certainly a causal effect on the feature, and (2) a reverse causation from feature to target is hard to argue.

We apply this scheme to the features of every task, after seeking advice from current research, governmental institutions and a medical practitioner. We describe the selection procedure for diagnosing diabetes in the following, and give more examples in Appendix A. Details on the feature selections of all tasks are provided in Appendix E.

2.1 Example: Variables in diabetes classification

The task is to classify whether a person is diagnosed with diabetes [Gardner et al., 2023]. The domains are defined by the preferred race of the individuals. We illustrate the feature grouping in Figure 3.

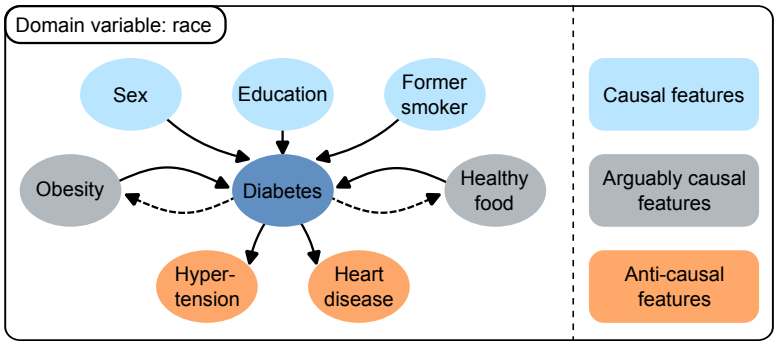


Figure 3: An example grouping for the task 'Diabetes'.

Causal features. Socio-economic status, in particular education level and former smoking, are widely acknowledged risk factors for diabetes [Brown et al., 2004, Agardh et al., 2011, Madduta et al., 2017, Centers for Disease Control and Prevention, 2024b]. Recent research in health care found evidence that an individual's sex impacts their diabetes diagnosis, e.g., pregnancies unmask pre-existing metabolic abnormalities in female individuals [Kautzky-Willer et al., 2023].[1] We also include marital status as a causal feature, as recent research showed that marital stress adversely affects the risk of developing diabetes [Whisman et al., 2014].

Arguably causal features. The individual's lifestyle, health and socio-economic status impacts their risk to develop diabetes, i.e., obesity, current smoking, healthy food, alcohol consumption, physical activities, mental health and utilization of health care services [Lindstrom et al., 2003, Brown et al., 2004, Engum, 2007, Baliunas et al., 2009, Agardh et al., 2011, Madduta et al., 2017, Centers for Disease Control and Prevention, 2024b, Klein et al., 2022, Centers for Disease Control and Prevention, 2024a]. At the same time, a person with diabetes is incentivized to improve their behavior to control their blood sugar and improve insulin sensitivity [Klein et al., 2004]. They are also more at risk to increase their weight due to the insulin therapy [McFarlane, 2009], experience distress [Centers for Disease Control and Prevention, 2024c], and have limited economic opportunities [American Diabetes Association, 2011]. Because of these bidirectional relationships, we regard features encoding these behaviors as arguably causal.

Anti-causal features. Researchers found evidence that diabetes increases the risk of hypertension, high blood cholesterol, coronary heart disease, myocardial infarction and strokes [Petrie et al., 2018, Schofield et al., 2016]. Due to treatment costs of diabetes, affected individuals are encouraged to obtain a health insurance [National Institute of Diabetes and Digestive and Kidney Diseases, 2019]. Therefore, we regard the current health care coverage as anti-causal to diabetes.

2.2 Tasks and datasets

We consider 16 classification tasks, listed in Table 1. The data is collected from a multitude of sources. We build on 14 classification tasks with natural domain shifts proposed in *TableShift*. We use the *TableShift* Python API to preprocess and transform raw public forms of the data.[2] In addition, we conduct experiments on two established classification tasks (MEPS, SIPP). Data preprocessing is adapted from Hardt and Kim [2023]. Further details on the tasks and their distribution shifts are in Appendix E.

2.3 Machine learning algorithms

In our experiments, we evaluate multiple machine learning algorithms. We list them in the following.

[1] There is an active debate in causal research whether non-manipulable variables like sex are proper causes [Holland, 2001, Pearl, 2018]. We acknowledge them as causes in our work.

[2] `https://tableshift.org/`

Table 1: Description of tasks, data sources and number of features in each selection. Details and licenses are provided in Appendix E.

Task	Data Source	#Features	#Arg. causal	#Causal	#Anti-causal
Food Stamps	ACS	28	25	12	-
Income	ACS	23	15	4	3
Public Coverage	ACS	19	16	8	-
Unemployment	ACS	26	21	11	3
Voting	ANES	54	36	8	-
Diabetes	BRFSS	25	17	4	6
Hypertension	BRFSS	18	14	5	2
College Scorecard	ED	118	34	11	-
ASSISTments	Kaggle	15	13	9	-
Stay in ICU	MIMIC-iii	7491	1445	5	-
Hospital Mortality	MIMIC-iii	7491	1445	5	-
Hospital Readmission	UCI	46	42	5	-
Childhood Lead	NHANES	7	6	5	-
Sepsis	PhysioNet	40	39	5	-
Utilization	MEPS	218	129	20	-
Poverty	SIPP	54	43	15	6

Baseline and tabular methods. We include tree ensemble methods: XGBoost [Chen and Guestrin, 2016], LightGBM [Ke et al., 2017] and histogram-based GBM. We also evaluate multilayer perceptrons (MLP) and state-of-the-art deep learning methods for tabular data: SAINT [Somepalli et al., 2021], TabTransformer [Huang et al., 2020], NODE [Popov et al., 2019], FT Transformer [Gorishniy et al., 2021] and tabular ResNet [Gorishniy et al., 2021].

Domain robustness and generalization methods. We consider distributionally robust optimization (DRO) [Levy et al., 2020], Group DRO [Sagawa et al., 2020] using domains and labels as groups, respectively, and the adversarial label robustness method by Zhang et al. [2021]. We also include Domain-Adversarial Neural Networks (DANN) [Ajakan et al., 2015], Deep CORAL [Sun and Saenko, 2016], Domain MixUp [Xu et al., 2020] and MMD [Li et al., 2018].

Causal methods. We assess Invariant Risk Minimization (IRM) [Arjovsky et al., 2019], Risk Extrapolation (REx) [Krueger et al., 2021], Information Bottleneck IRM (IB-IRM) [Ahuja et al., 2022], AND-Mask [Parascandolo et al., 2021] and CausIRL [Chevalley et al., 2022].

Domain generalization and causal methods require at least two training domains with a sufficient number of data points. This is provided in eight of our tasks. Detailed descriptions of the machine learning algorithms and hyperparameter choices are given in Appendix B and Gardner et al. [2023].

2.4 Experimental procedure

We conduct the following procedure for each task. First, we define up to four sets of features based on expert knowledge: all features, causal features, arguably causal features and anti-causal features. Second, we split the full dataset into in-domain set and out-of-domain set. We adopt the choice of domains from Gardner et al. [2023]. We have a train/test/validation split within the in-domain set, and a test/validation split within the out-of-domain set. For each feature set:

1. We apply the machine learning methods listed in Section 2.3. For each method:
   (a) We conduct a hyperparameter sweep using HyperOpt [Bergstra et al., 2013] on the in-domain validation data. A method is tuned for 50 trials. We exclusively train on the training set.
   (b) The trained classifiers are evaluated on in-domain and out-of-domain test set.
   (c) We select the best model according to their in-domain validation accuracy. This follows the selection procedure in previous work (e.g., [Gulrajani and Lopez-Paz, 2020, Gardner et al., 2023]). To ensure compatibility with *TableShift*, we add the best in-domain and

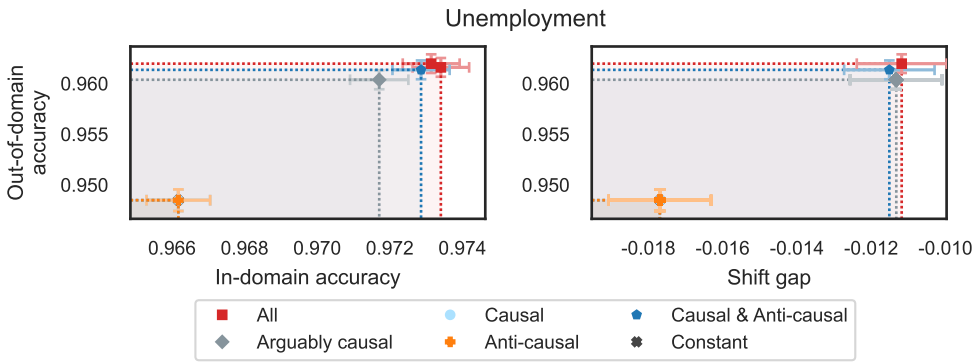


Figure 4: (Left) Pareto-frontiers of in-domain and out-of-domain accuracy by feature selection. (Right) Pareto-frontiers of shift gap and out-of-domain accuracy accomplished. Adding anti-causal features improves out-of-domain accuracy. Results of other tasks in Appendix C.

out-of-domain accuracy pair observed by [Gardner et al., 2023]. We restrict our further analysis to this selection.

2. We find the Pareto-set $\mathcal{P}$ of in-domain and out-of-domain accuracy pairs. We compute the shift gaps, and find the Pareto-set of shift gap and out-of-domain accuracy of the set $\mathcal{P}$.

We provide further details and illustrations of the individual steps in Appendix B.

3 Empirical results

In this section, we present and discuss the results of the experiments on all 16 tasks. A total of 42K models were trained for the main results and an additional 468K models for robustness tests. Our code is based on Gardner et al. [2023], Hardt and Kim [2023] and Gulrajani and Lopez-Paz [2020]. It is available at `https://github.com/socialfoundations/causal-features`.

In our experiments, we analyze the performance of feature selections based on domain-knowledge causal relations. A summary of the results is shown in Figure 1. Details on four representative tasks are given in Figure 2. The other tasks are in Appendix C. The accuracy results are presented along with 95% Clopper-Pearson intervals. They are the baseline for the approximate 95% confidence intervals of the shift gap. See Appendix B for the exact computation and justification of the confidence intervals.

In-domain and out-of-domain accuracy. Models trained on the whole feature set accomplish the highest in-domain and out-of-domain accuracy, up to error bars (16/16 tasks). The arguably causal features Pareto-dominate the causal features, up to error bars (16/16 tasks). Recall that arguably causal features are a superset of the causal features, and have considerably more features (Table 1). Models based on causal features often essentially predict the majority label (7/16 tasks).

Shift gap. The shift gap measures the absolute performance drop of the feature sets when employed out-of-domain. All features often experience a significantly smaller shift gap than causal features (7/16 tasks). The causal features solely surpass all features (within the error bounds) for the task 'Hospital Mortality' by predicting the majority label. In most cases, the shift gaps of all features and arguably causal features are indistinguishable (15/16 tasks).

Anti-causal features. In five tasks, we have features that we regard as anti-causal. Results are shown in Figure 4 and Appendix C.2. The anti-causal features do not perform significantly different from the constant predictor in-domain (5/5 tasks). However, they sometimes perform extremely poor out-of-domain (2/5 tasks). It is therefore astounding that the out-of-domain performance of the (arguably) causal features is improved by adding anti-causal features (5/5 tasks).

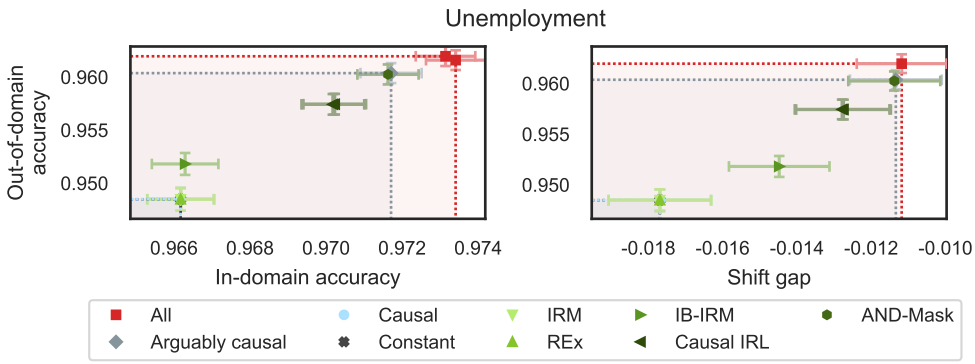


Figure 5: (Left) Pareto-frontiers of in-domain and out-of-domain accuracy of causal methods and domain-knowledge features selection. (Right) Pareto-frontiers of shift gap and out-of-domain accuracy attained. The performance of the causal methods interpolates between the performance of the causal and/or arguably causal features. Results of remaining tasks are in Appendix C.

Causal machine learning methods. We restrict ourselves to the Pareto-set of the standard models for each feature set and compare them to causal methods.[3] We showcase a representative performance in Figure 5. Details are in Appendix C.3. The causal methods do not improve upon the arguably causal features trained on standard models (8/8 tasks). In fact, their performance typically spans between the causal features and arguably causal methods trained on standard models. The in-domain and out-of-domain accuracy is even indistinguishable from the causal selections in multiple cases (IRM: 3/8, REx: 4/8, IB-IRM: 2/8, CausIRL: 5/8, AND-Mask: 5/8 tasks). Possible explanations are: (1) the causal methods manage to extract a causal representation of the features similar to our selection of causal and/or arguably causal features; or (2) it is an artifact of having low predictive power.

Causal discovery algorithms. We apply invariant causal prediction (ICP) [Peters et al., 2016] and classic causal discovery algorithm, Peter-Clark (PC) algorithm [Spirtes et al., 2000] and Fast causal inference (FCI) algorithm [Spirtes et al., 1995], to our tasks. See Appendix C.4 for the results. The algorithms rarely outputs any causal parents (ICP: 1/6, PC: 4/14, FCI: 1/14 tasks). When they do, they select very few features as causal parents. Some of them are features we also regard as causal based on domain knowledge, others anti-causal or without causal relations to the target. For example, PC outputs an individual's occupation and the number of weeks worked in the last 12 months as causal parents of unemployment. While we agree that the occupation has a causal influence on unemployment, we view the amount of weeks an individual worked as a *result* of their unemployment rather than the *reason*. The causal parents estimated by the causal discovery algorithms often perform similar to the causal features though (ICP: 1/1, PC: 1/4, FCI: 1/1 tasks). Note that their performance is always Pareto-dominated by our arguably causal features (ICP: 1/1, PC: 4/4, FCI: 1/1 tasks). Therefore, whichever feature selection one choose to believe, ours or causal discovery algorithms', one never improves upon the whole feature set.

Robustness tests. Results are in Appendix C.1, C.5 and C.6. We test whether our conclusions are sensitive to misclassifying one feature. Therefore, we form subsets of the set of causal features by removing one feature at a time. The test subsets do not achieve higher out-of-domain accuracy than using all features, with one exception in the task 'ASSISTments'. We find that the supersets of arguably causal features with one additional features obtain similar or better out-of-domain accuracy. We randomly sample 500 feature subsets for each task and check whether any subset significantly outperforms the whole feature set. None of the sampled subset does, with few outliers in the task 'ASSISTments'. We consider the divergent task in detail. The task is about predicting whether a student answers a question correct. Surprisingly, all outperforming random subsets and the subset from the misclassification tests coincide in one regard: *missing* the feature encoding the tested skill, e.g., rounding. We encourage further work to explain this oddity, as the tested skill of a task

[3]We refer machine learning methods that are not explicitly causally motivated as *standard*: baseline methods, tabular methods, domain robustness methods and non-causal domain generalization methods.

clearly has a causal influence. We also provide insights into which non-causal features improve the out-of-domain performance, and discuss potential explanations.
Our findings remain valid when using balanced accuracy as a metric.

4 Discussion and limitations

Our findings may not come as a surprise to everyone. Unlike causal machine learning researchers, social scientists generally see no reason to believe in the universality of causal relationships. For example, smaller classroom sizes may cause better teaching outcomes in Tennessee [Mosteller, 1995], but much less so in California [Jepsen and Rivkin, 2009]. Such variation is the rule rather than the exception. Indeed, philosopher of science Cartwright [1999, 2007] argued that causal regularities are often more narrowly scoped than commonly held.

Our study mirrors these robust facts in a machine learning context. In the many common tabular datasets we consider, we find no evidence that causal predictors have greater external validity than their conventional counterparts. If the goal is to generalize to new domains in these datasets, our findings suggest we might as well train the best possible model on all available features. The one exception to the rule we found is the case of the *skill* variable in the Kaggle ASSISTments task. It appears as though removing this variable increases out-of-domain generalization. Curiously, the variable is also almost certainly one of the better examples of a *causal* variable in our study. Removing it therefore gives no advantage to causal predictors. For all tasks, we used available research and our own knowledge to classify variables as causal and arguably causal. We likely made some mistakes in this classification. This is why we extended our study with extensive robustness checks that confirm our findings. In addition, we did not find any relief in state-of-the-art causal methods, or causal discovery algorithms.

Demonstrating the utility of causal methods therefore likely requires other benchmark datasets than the ones currently available. We consider this a promising avenue for future work that derives further motivation from our work. We point to two classification tasks, where recent research suggests that causal prediction methods have utility for better domain generalization [Schulam and Saria, 2017, Subbaswamy and Saria, 2019]: predicting the probability pneumonia mortality outside the hospitals [Cooper et al., 1997, Caruana et al., 2015], and hospital mortality across changes in the clinical information system [Nestor et al., 2019]. We refer to Appendix A for details. Another direction for future research is to evaluate to which extent our findings generalize to other applications and data modalities. Recent advances in causal machine learning suggest, for example, promising results in real-world image datasets for classifying wild animals (Terra Incognita) and urban vs. non-urban examples (Spurious PACS), see [Salaudeen and Koyejo, 2024].

In light of our results, it's worth finding theoretical explanations for why using all features, regardless of causality, has the best performance in typical tabular datasets. In this vein, Rosenfeld et al. [2021a] point to settings where risk minimization is the right thing to do in theory. We seed the search for additional theoretical explanations with a simple observation: If all domains are positive reweightings of one another, then the Bayes optimal predictor with respect to classification error in one domain is also Bayes optimal in any other domain. Standard models, such as gradient boosting or random forests, often achieve near optimal performance on tabular data with a relatively small number of features. In such cases, our simple observation applies and motivates a common sense heuristic: Do the best you can to approximate the optimal predictor on all available features.

Acknowledgments and Disclosure of Funding

We're indebted to Sabrina Doleschel for her help in classifying the arguably causal features in task 'Length of Stay' and 'Hospital Mortality' based on MIMIC-III. We thank André Cruz, Ricardo Dominguez-Olmedo, Florian Dorner, Mila Gorecki, Markus Kalisch, Nicolai Meinshausen, and Olawale Salaudeen for invaluable feedback on an earlier version of this paper. In addition, we thank Josh Gardner, Francesco Montagna and Ricardo J. Sandoval for sharing their code with us and being open for follow-up discussions.

This project has received funding from the Max Planck ETH Center for Learning Systems (CLS).

References

E. Agardh, P. Allebeck, J. Hallqvist, T. Moradi, and A. Sidorchuk. Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. *International Journal of Epidemiology*, 40(3):804–818, 02 2011. URL http://doi.org/10.1093/ije/dyr029.

Agency for Healthcare Research and Quality. Medical expenditure panel survey (MEPS), 2019.

K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. 2022. URL http://doi.org/10.48550/arXiv.2106.06607.

H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. Domain-adversarial neural networks. 2015. URL http://doi.org/10.48550/arXiv.1412.4446.

R. Akee, M. R. Jones, and S. R. Porter. Race matters: Income shares, income inequality, and income mobility for all U.S. races. Working Paper 23733, National Bureau of Economic Research, 2017. URL http://doi.org/10.3386/w23733.

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

M. Ameri, L. Schur, M. Adya, S. Bentley, P. McKay, and D. Kruse. The disability employment puzzle: A field experiment on employer hiring behavior. Working Paper 21560, National Bureau of Economic Research, 2015. URL http://doi.org/10.3386/w21560.

American Diabetes Association. Diabetes and employment. *Diabetes Care*, 34:S82–S86, 2011. URL http://doi.org/10.2337/dc11-S082.

American National Election Studies. ANES time series cumulative data file, 1948-2020, 2020.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

T. Averbuch, M. Mohamed, S. Islam, E. Defilippis, K. Breathett, M. Alkhouli, E. Michos, G. Martin, E. Kontopantelis, M. Mamas, and H. Van Spall. The association between socioeconomic status, sex, race / ethnicity and in-hospital mortality among patients hospitalized for heart failure. *Journal of Cardiac Failure*, 28(5):697–709, 2022. URL http://doi.org/10.1016/j.cardfail.2021.09.012.

R. S. Baker, S. K. D'Mello, M. T. Rodrigo, and A. C. Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010. ISSN http://doi.org/10.1-5819. URL http://doi.org/10.1016/j.ijhcs.2009.12.003.

D. O. Baliunas, B. J. Taylor, H. Irving, M. Roerecke, J. Patra, S. Mohapatra, and J. Rehm. Alcohol as a risk factor for type 2 diabetes: A systematic review and meta-analysis. *Diabetes Care*, 32(11):2123–2132, 11 2009. URL http://doi.org/10.2337/dc09-0227.

C. N. Bell, R. J. Thorpe, and T. A. LaVeist. Race/ethnicity and hypertension: The role of social support. *American Journal of Hypertension*, 23(5):534–540, 2010. URL http://doi.org/10.1038/ajh.2010.28.

J. Bendor, D. Diermeier, and M. Ting. A behavioral model of turnout. *The American Political Science Review*, 97(2):261–280, 2003. URL http://www.jstor.org/stable/3118208.

J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, 2013. PMLR. URL https://proceedings.mlr.press/v28/bergstra13.html.

F. D. Blau and L. M. Kahn. The gender wage gap: Extent, trends, and explanations. Working Paper 21913, National Bureau of Economic Research, 2016. URL http://doi.org/10.3386/w21913.

S. Bonaccio, C. Connelly, I. Gellatly, A. Jetha, A. Jetha, and K. A. M. Ginis. The participation of people with disabilities in the workplace across the employment cycle: Employer concerns and research evidence. *Journal of Business and Psychology*, 35:135–158, 2019. URL http://doi.org/10.1007/s10869-018-9602-5.

C. Bosquet and H. G. Overman. Why does birthplace matter so much? *Journal of Urban Economics*, 110:26–34, 2019. URL http://doi.org/10.1016/j.jue.2019.01.003.

A. F. Brown, S. L. Ettner, J. Piette, M. Weinberger, E. Gregg, M. F. Shapiro, A. J. Karter, M. Safford, B. Waitzfelder, P. A. Prata, and G. L. Beckles. Socioeconomic position and health among persons with diabetes mellitus: A conceptual framework and review of the literature. *Epidemiologic Reviews*, 26(1):63–77, 07 2004. URL http://doi.org/10.1093/epirev/mxh002.

R. E. Bulanda and J. R. Bulanda. *Six: The Problem of Unpaid Parental Leave*, pages 53–62. Policy Press, 2020. URL http://doi.org/10.51952/9781447354611.ch006.

Bureau of Economic Analysis. Personal income by state, 2024. URL https://www.bea.gov/data/income-saving/personal-income-by-state.

D. Card. The causal effect of education on earnings. volume 3 of *Handbook of Labor Economics*, pages 1801–1863. Elsevier, 1999. URL http://doi.org/10.1016/S1573-4463(99)03011-4.

N. Cartwright. *The dappled world: A study of the boundaries of science*. Cambridge University Press, 1999.

N. Cartwright. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press, 2007.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730. Association for Computing Machinery, 2015. URL http://doi.org/10.1145/2783258.2788613.

Centers for Disease Control and Prevention. National health and nutrition examination survey questionnaire, examination protocol, and laboratory protocol (1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017), 2017.

Centers for Disease Control and Prevention. BRFSS Survey Data (2015, 2017, 2019, 2021), 2021.

Centers for Disease Control and Prevention. Diabetes risk factors, 2024a. URL https://www.cdc.gov/diabetes/risk-factors.

Centers for Disease Control and Prevention. Smoking and diabetes, 2024b. URL https://www.cdc.gov/diabetes/risk-factors/diabetes-and-smoking.html.

Centers for Disease Control and Prevention. Diabetes and mental health, 2024c. URL https://www.cdc.gov/diabetes/living-with/mental-health.html.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 785–794, 2016.

Y. Chen, K. Zhou, Y. Bian, B. Xie, B. Wu, Y. Zhang, M. KAILI, H. Yang, P. Zhao, B. Han, and J. Cheng. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=esFxSb_0pSL.

M. Chevalley, C. Bunne, A. Krause, and S. Bauer. Invariant causal mechanisms through distribution matching. 2022. URL http://doi.org/10.48550/arXiv.2206.11646.

S. Choi and A. Valladares-Esteban. The marriage unemployment gap. *The B.E. Journal of Macroeconomics*, 18(1):20160060, 2018. URL http://doi.org/10.1515/bejm-2016-0060.

J. Clore, K. Cios, J. DeShazo, and B. Strack. Diabetes 130-US hospitals for years 1999-2008. UCI Machine Learning Repository, 2014. URL http://doi.org/10.24432/C5230J.

G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9(2):107–138, 1997. URL http://doi.org/10.1016/S0933-3657(96)00367-3.

M. M. L. Crepaz. The impact of party polarization and postmaterialism on voter turnout. *European Journal of Political Research*, 18(2):183–205, 1990. URL http://doi.org/10.1111/j.1475-6765.1990.tb00228.x.

G. F. De Jong and A. B. Madamba. A double disadvantage? Minority group, immigrant status, and underemployment in the united states. *Social Science Quarterly*, 82(1):117–130, 2001. URL http://doi.org/10.1111/0038-4941.00011.

F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, NeurIPS 2021, pages 6478–6490, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

K. S. Dorans, K. T. Mills, Y. Liu, and J. He. Trends in prevalence and control of hypertension according to the 2017 American college of cardiology/American heart association (acc/aha) guideline. *Journal of the American Heart Association*, 7(11):e008888, 2018. URL http://doi.org/10.1161/JAHA.118.008888.

A. Engum. The role of depression and anxiety in onset of diabetes in a large population-based study. *Journal of Psychosomatic Research*, 62(1):31–38, 2007. URL http://doi.org/10.1016/j.jpsychores.2006.07.009.

J. Ermisch and M. Francesconi. The effect of parental employment on child schooling. *Journal of Applied Econometrics*, 28(5):796–822, 2013. URL http://doi.org/10.1002/jae.2260.

M. Feng, N. Heffernan, , and K. Koedinger. Addressing the assessment challenge in an intelligent tutoring system that tutors as it assesses. *The Journal of User Modeling and User-Adapted Interaction*, 19:243–266, 2009.

J. Gardner, Z. Popovic, and L. Schmidt. Benchmarking distribution shift in tabular data with tableshift. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, NeurIPS 2023, 2023. URL https://proceedings.neurips.cc/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

T. A. Gerace, J. Hollis, J. K. Ockene, and K. Svendsen. Smoking cessation and change in diastolic blood pressure, body weight, and plasma lipids, 1991.

A. L. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. URL http://doi.org/10.1161/01.cir.101.23.e215.

Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943. Curran Associates, Inc., 2021.

I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. 2020. URL http://doi.org/10.48550/arXiv.2007.01434.

T. Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115, 1944.

D. Halvarsson, M. Korpi, and K. Wennberg. Entrepreneurship and income inequality. *Journal of Economic Behavior & Organization*, 145:275–293, 2018. URL http://doi.org/10.1016/j.jebo.2017.11.003.

M. Hardt and M. P. Kim. Is your model predicting the past? In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. URL http://doi.org/10.1145/3617694.3623225.

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

P. W. Holland. The false linking of race and causality: lessons from standardized testing. *Race and Society*, 4(2):219–233, 2001. URL http://doi.org/10.1016/S1090-9524(03)00011-1.

K. D. Hoover. The logic of causal inference: Econometrics and the conditional analysis of causation. *Economics & Philosophy*, 6(2):207–234, 1990.

X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. 2020. URL http://doi.org/10.48550/arXiv.2012.06678.

C. Jepsen and S. Rivkin. Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of human resources*, 44(1):223–250, 2009.

Y. Jiang and V. Veitch. Invariant and transportable representations for anti-causal domain shifts. 2022. URL http://doi.org/10.48550/arXiv.2207.01603.

A. E. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva. Causal machine learning: A survey and open problems. 2022. URL http://doi.org/10.48550/arXiv.2206.15475.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012. URL http://doi.org/10.18637/jss.v047.i11.

J. S. Kaufman, L. Dolman, D. Rushani, and R. S. Cooper. The contribution of genomic research to explaining racial disparities in cardiovascular disease: A systematic review. *American Journal of Epidemiology*, 181(7):464–472, 03 2015. URL http://doi.org/10.1093/aje/kwu319.

J. N. Kaur, E. Kiciman, and A. Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. 2022. URL http://doi.org/10.48550/arXiv.2206.07837.

A. Kautzky-Willer, M. Leutner, and J. Harreiter. Sex differences in type 2 diabetes. *Diabetologia*, 66 (6):986–1002, 2023. URL http://doi.org/10.1007/s00125-023-05891-x.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3149–3157. Curran Associates Inc., 2017. URL http://doi.org/10.5555/3294996.3295074.

S. Klein, N. F. Sheard, X. Pi-Sunyer, A. Daly, J. Wylie-Rosett, K. Kulkarni, and N. G. Clark. Weight management through lifestyle modification for the prevention and management of type 2 diabetes: rationale and strategies. *The American journal of clinical nutrition*, 80(2):257–263, 2004.

S. Klein, A. Gastaldelli, H. Yki-Järvinen, and P. E. Scherer. Why does obesity cause diabetes? *Cell Metabolism*, 34(1):11–20, 2022. URL http://doi.org/10.1016/j.cmet.2021.12.012.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

J. E. Leighley and J. Nagler. *Who Votes Now?: Demographics, Issues, Inequality, and Turnout in the United States*. 2014. URL http://doi.org/10.1515/9781400848621.

B. Leng, Y. Jin, G. Li, L. Chen, and N. Jin. Socioeconomic status and hypertension: a meta-analysis. *Journal of Hypertension*, 33(2):221–229, 2015.

D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8847–8860. Curran Associates, Inc., 2020.

H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. URL http://doi.org/10.1109/CVPR.2018.00566.

J. Lindstrom, A. Louheranta, M. Mannelin, M. Rastas, V. Salminen, J. Eriksson, M. Uusitupa, J. Tuomilehto, and F. D. P. S. Group. The finnish diabetes prevention study (dps) lifestyle intervention and 3-year results on diet and physical activity. *Diabetes care*, 26(12):3230–3236, 2003.

D. S. Loughran. *Why Is Veteran Unemployment So High?* RAND Corporation, Santa Monica, CA, 2014. URL http://doi.org/10.7249/RR284.

J. Madduta, E. Anderson-Baucum, and C. Evans-Molina. Smoking and the risk of type 2 diabetes. *Translational research*, 184:101–107, 2017. URL http://doi.org/10.1016/j.trsl.2017.02.004.

C. M. McEniery, I. B. Wilkinson, and A. P. Avolio. Age, hypertension and arterial function. *Clinical and Experimental Pharmacology and Physiology*, 34(7):665–671, 2007. URL http://doi.org/10.1111/j.1440-1681.2007.04657.x.

S. McFarlane. Insulin therapy and type 2 diabetes: management of weight gain. *Journal of clinical hypertension*, 11(10):601–607, 2009. URL http://doi.org/10.1111/j.1751-7176.2009.00063.x.

N. Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10, 2018. URL http://doi.org/10.1109/DSW.2018.8439889.

J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7721–7735. PMLR, 18-24 Jul 2021. URL https://proceedings.mlr.press/v139/miller21b.html.

K. T. Mills, A. Stefanescu, and J. He. The global epidemiology of hypertension. *Nature reviews*, 16 (4):223–237, 2020. URL http://doi.org/10.1038/s41581-019-0244-2.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In M. van der Schaar, C. Zhang, and D. Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 752–771. PMLR, 2023.

F. Montagna, A. Mastakouri, E. Eulig, N. Noceti, L. Rosasco, D. Janzing, B. Aragam, and F. Locatello. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.

F. Mosteller. The tennessee study of class size in the early school grades. *The future of children*, pages 113–127, 1995.

National Centers for Environmental Information. U.S. Census Divisions, 2024. URL https://www.ncei.noaa.gov/access/monitoring/reference-maps/us-census-divisions.

National Institute of Diabetes and Digestive and Kidney Diseases. Financial help for diabetes care, 2019. URL https://www.niddk.nih.gov/health-information/diabetes/financial-help-diabetes-care.

B. Nestor, M. B. A. McDermott, W. Boag, G. Berner, T. Naumann, M. C. Hughes, A. Goldenberg, and M. Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 381–405. PMLR, 2019. URL https://proceedings.mlr.press/v106/nestor19a.html.

S. Pandey, A. Kalaria, K. D. Jhaveri, S. M. Herrmann, and A. S. Kim. Management of hypertension in patients with cancer: challenges and considerations. *Clinical Kidney Journal*, 16(12):2336–2348, 2023. URL http://doi.org/10.1093/ckj/sfad195.

G. Parascandolo, A. Neitz, A. ORVIETO, L. Gresele, and B. Schölkopf. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hb1sDDSLbV.

J. H. Park. The earnings of immigrants in the United States: the effect of English-speaking ability. *American Journal of Economics and Sociology*, pages 43–56, 1999.

J. Pearl. *Causality*. Cambridge University Press, 2009.

J. Pearl. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference*, 6(2):20182001, 2018. URL http://doi.org/10.1515/jci-2018-2001.

J. Pearl and E. Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 451–482. 2022.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. URL https://mitpress.mit.edu/books/elements-causal-inference.

J. R. Petrie, T. J. Guzik, and R. M. Touyz. Diabetes, hypertension, and cardiovascular disease: Clinical insights and vascular mechanisms. *The Canadian journal of cardiology*, 34(5):575–584, 2018. URL http://doi.org/10.1016/j.cjca.2017.12.005.

S. Popov, S. Morozov, and A. Babenko. Neural oblivious decision ensembles for deep learning on tabular data. 2019. URL http://doi.org/10.48550/arXiv.1909.06312.

R. Putnam. *Bowling Alone: The Collapse and Revival of American Community*. A Touchstone book. Simon & Schuster, 2000. URL https://books.google.de/books?id=rd2ibodep7UC.

L. A. Ramirez and J. C. Sullivan. Sex differences in hypertension: Where we have been and where we are going. *American Journal of Hypertension*, 31(12):1247–1254, 2018. URL http://doi.org/10.1093/ajh/hpy148.

M. A. Reyna, C. Josef, R. Jeter, S. Shashikumar, B. Moody, M. B. Westover, A. Sharma, S. Nemati, and G. D. Clifford. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019 (version 1.0.0). *PhysioNet*, 2019. URL http://doi.org/10.13026/v64v-d857.

M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48(2):210–217, 2020. URL http://doi.org/10.1097/CCM.0000000000004145.

J. C. Rogowski. Electoral choice, ideological conflict, and political participation. *American Journal of Political Science*, 58(2):479–494, 2014. URL http://doi.org/10.1111/ajps.12059.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. URL http://jmlr.org/papers/v19/16-432.html.

P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18741–18753. PMLR, 2022.

E. Rosenfeld, P. Ravikumar, and A. Risteski. An online learning approach to interpolation and extrapolation in domain generalization. 2021a. URL http://doi.org/10.48550/arXiv.2102.13128.

E. Rosenfeld, P. Ravikumar, and A. Risteski. The risks of invariant risk minimization. 2021b. URL http://doi.org/10.48550/arXiv.2010.05761.

S. J. Rosenstone. Economic adversity and voter turnout. *American Journal of Political Science*, 26 (1):25–46, 1982. URL http://www.jstor.org/stable/2110837.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. 2020. URL http://doi.org/10.48550/arXiv.1801.06229.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. 2020. URL http://doi.org/10.48550/arXiv.1911.08731.

O. Salaudeen and S. Koyejo. Causally inspired regularization enables domain general representations. In *International Conference on Artificial Intelligence and Statistics*, pages 3124–3132. PMLR, 2024.

E. Salter. Deciding for a child: a comprehensive analysis of the best interest standard. *Theor Med Bioeth*, 33:179–198, 2012. URL http://doi.org/10.1007/s11017-012-9219-z.

C. Schoen, K. Davis, C. DesRoches, K. Donelan, and R. Blendon. Health insurance markets and income inequality: findings from an international health policy survey. *Health Policy*, 51(2):67–85, 2000. ISSN 0168-8510. URL http://doi.org/10.1016/S0168-8510(99)00084-6.

J. D. Schofield, Y. Liu, P. Rao-Balakrishna, R. A. Malik, and H. Soran. Diabetes dyslipidemia. *Diabetes therapy*, 7(2):203–219, 2016. URL http://doi.org/10.1007/s13300-016-0167-x.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML 2012, pages 459–466, 2012. URL http://doi.org/10.5555/3042573.3042635.

P. Schulam and S. Saria. Reliable decision support using counterfactual models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/299a23a2291e2126b91d54f3601ec162-Paper.pdf.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, http://doi.org/10.(5):612–634, 2021. URL http://doi.org/10.1109/JPROC.2021.3058954.

A. Sionakidis, L. McCallum, and S. Padmanabhan. Unravelling the tangled web of hypertension and cancer. *Clinical Science*, 135(13):1609–1625, 2021. URL http://doi.org/10.1042/CS20200307.

V. Skirbekk. Age and individual productivity: A literature survey. *Vienna Yearbook of Population Research*, 2:133–153, 2004.

S. Soffer, E. Zimlichman, B. S. Glicksberg, O. Efros, M. A. Levin, R. Freeman, D. L. Reich, and E. Klang. Obesity as a mortality risk factor in the medical ward: a case control study. *BMC Endocrine Disorders*, 22(1):1–8, 2022.

G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. 2021. URL http://doi.org/10.48550/arXiv.2106.01342.

R. M. Sondheimer and D. P. Green. Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science*, 54(1):174–189, 2010. URL http://doi.org/10.1111/j.1540-5907.2009.00425.x.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. URL http://doi.org/10.5555/2074158.2074215.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014, 2014.

A. Subbaswamy and S. Saria. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In R. Silva, A. Globerson, and A. Globerson, editors, *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, volume 2, pages 947–957. Association For Uncertainty in Artificial Intelligence (AUAI), Jan. 2018.

A. Subbaswamy and S. Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, 11 2019. URL http://doi.org/10.1093/biostatistics/kxz041.

A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. 2019. URL http://doi.org/10.48550/arXiv.1812.04597.

A. Subbaswamy, B. Chen, and S. Saria. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, 10(1):64–89, 2022. URL http://doi.org/10.1515/jci-2021-0042.

B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. 2016. URL http://doi.org/10.48550/arXiv.1607.01719.

B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. 2015. URL http://doi.org/10.48550/arXiv.1511.05547.

P. Taubman. Role of parental income in educational attainment. *The American Economic Review*, 79 (2):57–61, 1989. ISSN 00028282. URL http://www.jstor.org/stable/1827730.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

U.S. Army. Take charge of your present and future, 2024. URL https://www.goarmy.com/benefits.html.

U.S. Bureau of Labor Statistics. Average weeks unemployed [lnu03008275], 2024. URL https://fred.stlouisfed.org/series/LNU03008275.

U.S. Census Bureau. Survey of income and program participation (SIPP), 2014.

U.S. Census Bureau. American community survey (ACS), 2018.

U.S. Census Bureau. Medical expenditure panel survey - insurance component shows 86% of private-sector employees worked for establishments that offered health insurance, 2024. URL https://www.census.gov/library/stories/2024/02/health-care-costs.html.

U.S. Centers for Medicare and Medicaid Services. Medicaid, 2024a. URL https://www.medicaid.gov/.

U.S. Centers for Medicare and Medicaid Services. Medicare, 2024b. URL https://www.medicare.gov/.

U.S. Citizenship and Immigration Services. Should I consider U.S. citizenship?, 2024. URL https://www.uscis.gov/citizenship/learn-about-citizenship/should-i-consider-us-citizenship.

U.S. Department of Education. College scorecard, 2023.

U.S. Department of Labor. Employment and earnings by occupation, 2024. URL https://www.dol.gov/agencies/wb/data/occupations.

U.S. General Services Administration. Register to vote, 2024. URL https://vote.gov/.

U.S. Office of Foreign Labor Certification. Seasonal jobs, 2024. URL https://seasonaljobs.dol.gov/.

A. Virdis, C. Giannarelli, M. F. Neves, S. Taddei, and L. Ghiadoni. Cigarette smoking and hypertension. *Current pharmaceutical design*, 16(3):2518–2525, 2010. URL http://doi.org/10.2174/138161210792062920.

M. Walicka, M. Puzianowska-Kuznicka, M. Chlebus, A. Śliwczyński, M. Brzozowska, D. Rutkowski, L. Kania, M. Czech, A. Jacyna, and E. Franek. Relationship between age and in-hospital mortality during 15,345,025 non-surgical hospitalizations. *Archives of Medical Science*, 17(1):40–46, 2021. URL http://doi.org/10.5114/aoms/89768.

H. Wang, H. Si, B. Li, and H. Zhao. Provable domain generalization via invariant-feature subspace recovery. 2022a. URL http://doi.org/10.48550/arXiv.2201.12919.

J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu. Generalizing to unseen domains: A survey on domain generalization. 2022b. URL http://doi.org/10.48550/arXiv.2103.03097.

S. Wang, M. B. A. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 222–235. Association for Computing Machinery, 2020a. URL http://doi.org/10.1145/3368555.3384469.

S. Wang, M. B. A. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, ACM CHIL '20. ACM, 2020b. URL http://doi.org/10.1145/3368555.3384469.

M. A. Whisman, A. Li, D. A. Sbarra, and C. L. Raison. Marital quality and diabetes: results from the health and retirement study. *Health Psychology*, 33(8):832, 2014.

M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang. Adversarial domain adaptation with domain mixup. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6502–6509, 2020. URL http://doi.org/10.1609/aaai.v34i04.6123.

J. Zhang, A. Menon, A. Veit, S. Bhojanapalli, S. Kumar, and S. Sra. Coping with label shift via distributionally robust optimisation. 2021. URL http://doi.org/10.48550/arXiv.2010.12230.

Supplemental material

We provide detailed descriptions and results to supplement our main paper, as listed in the following.

List of Appendices

A Examples

In this section, we explain our feature selection for multiple examples. We also discuss examples of datasets that motivate incorporating causal theory to enhance predictions.

A.1 Selection into causal, arguably causal and anti-causal features

We choose seven representatives tasks to illustrate how we select causal, arguably causal and anti-causal features. These tasks include predicting: economic outcomes (income level and unemployment), political decisions (whether an individual vote in the presidential election), health diagnoses (diabetes, hypertension and hospital mortality) and educational achievements (correctly solved problems in an online learning program). We refer to Section 2.1 for the task 'Diabetes'.

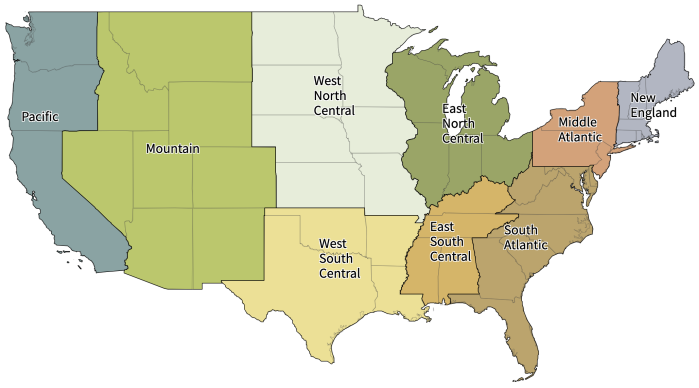


Figure 6: US Census Divisions. Source: National Centers for Environmental Information [2024]

A.1.1 Income

The task is to classify whether a person has a low or high income level [Gardner et al., 2023]. The domains are defined by the U.S. Census Divisions, with New England being set as out-of-domain. See Figure 6. We provide a list of all features in Appendix E.1.2.

Causal features. Skirbekk [2004] showed that an individual's age affects their productivity, that is, their job performance, and therefore, their income. Marginalized groups, such as women and people of color, face discrimination in pursuing higher income, apparent in gender and race wage gaps [Blau and Kahn, 2016, Akee et al., 2017]. Therefore, we include the self-reported age, gender and race as causal features. Bosquet and Overman [2019] also found evidence that the place of birth influences an individual's later income.

Arguably causal features. Personal income varies across U.S. states due to different economic situations [Bureau of Economic Analysis, 2024]. These diverse economic opportunities, on the other hand, lead to internal migration across states. Hence, we regard the individual's current state of residence merely as arguably causal. While educational attainment and the ability to speak English enable higher-paying jobs [Card, 1999, Park, 1999], a certain level of income is needed to pay for college tuition or language courses [Taubman, 1989]. Different occupations differ drastically in their annual earnings [U.S. Department of Labor, 2024]. The specific work habits, e.g. number of weeks worked and usual hours worked per week, naturally affect the individual's earnings and thus income. Then again, the choice of occupation and work habits may stem from a certain level of income from other sources, e.g. investments or lack thereof [Halvarsson et al., 2018]. Therefore, all work-related features are regarded as arguably causal. Giving birth to a child usually leads to a short-timed drop in income due to (unpaid) parental leave or child care [Bulanda and Bulanda, 2020]. A certain level of income may, however, be a consideration for some people when deciding on a child [Salter, 2012]. Citizenship enables individuals for governmental jobs and certain social benefits, alluring people to obtain U.S. citizenship [U.S. Citizenship and Immigration Services, 2024].

Anti-causal features. Insurance purchased directly from an insurance company requires a certain level of income to cover the regular payments [Schoen et al., 2000]. On the other hand, low income is a requirement to be able to apply for Medicaid and other government assistance plans [U.S. Centers for Medicare and Medicaid Services, 2024a]. Hence, we view these insurance types as features that are anti-causally related to income. A low level of income may necessitate a person to look for work, if they haven't any yet.

Other features. We don't see any obvious direct causal link between a person's income and their marital status. Insurance through an employer or union and Medicare are benefits not tied to income, but rather the person's employer or age and medical condition [U.S. Census Bureau, 2024, U.S. Centers for Medicare and Medicaid Services, 2024b].

A.1.2 Unemployment

The task is to classify whether a person is unemployed [Gardner et al., 2023]. The domains are defined by the education level, with no high school diploma being out-of-domain. We provide a list of all features in Appendix E.1.4.

Causal features. Some people are unable to work due to their disability. Even if they are capable to work, many employers are still unwilling to (continue to) employ them [Ameri et al., 2015, Bonaccio et al., 2019]. Hence, we regard features noting the self-reported disabilities as causal. Moreover, immigrant workers also face initial disadvantages in labor force assimilation [De Jong and Madamba, 2001]. The immigrant status is encoded as the place of birth in our features. We also view age, sex, race and ancestry as causal. The same arguments apply as in Section A.1.1.

Arguably causal features. Some occupations are mainly seasonal, e.g., working on farms, in landscape or construction [U.S. Office of Foreign Labor Certification, 2024], and therefore, may lead to regular short-term unemployment. On the other hand, being unemployed may necessitate an individual to take on training and change their chosen occupation. Being unemployed and looking for work may motivate an individual to consider joining the armed forces [U.S. Army, 2024]. Later on however, veterans are less likely to be employed due to poor health, employer discrimination, or skill mismatch [Loughran, 2014]. Hence, unemployment and occupation/military service are tangled together in a complex way, which is why we regard the feature encoding them as arguably causal. The ability to speak English is a requirement in some jobs. Conversely, an individual may learn English naturally by interacting with their co-workers. The family situation, described by the marital status and the employment status of the parents, may (indirectly) impact the individual's unemployment. Choi and Valladares-Esteban [2018] show that single workers face higher job losing probabilities than married ones, and multiple studies establish that the employment status of an individual's parents impacts the child's attainments [Taubman, 1989, Ermisch and Francesconi, 2013]. Similar arguments as in Section A.1.1 apply for citizenship, current state, mobility status and giving birth to a child.

Anti-causal features. As people are unemployed on average for around 15 to 25 weeks in the U.S. [U.S. Bureau of Labor Statistics, 2024], unemployment directly impacts the number of weeks worked during the past 12 months, the usual hours worked per week and whether a person worked last week.

A.1.3 Voting

The task is to classify whether a person voted in the U.S. presidential election [Gardner et al., 2023]. The domains are defined by U.S. Census Regions, with South set as out-of-domain. South consists of the U.S. Census divisions West South Central, East South Central and South Atlantic. See Figure 6. We provide a list of all features in Appendix E.2.1.

Causal features. All states except North Dakota require that a person register before voting in an election [U.S. General Services Administration, 2024], which is one of our features. Leighley and Nagler [2014] discuss in detail the difference in voting behavior between demographic groups, e.g., defined by age, gender, race/ethnicity, and state. Hence, we again view the demographic features as causal. There is also evidence that education and occupation influence the decision to vote [Sondheimer and Green, 2010, Rosenstone, 1982]. The current social climate and ideological conflict between competing electoral options also affects voter turnout, encoded by the election year [Rogowski, 2014, Putnam, 2000].

Arguably causal features. Participating in politics, being interested in the election, or at least being confronted with the election via media may strengthen a person's resolve to vote. As deciding to vote can not be explained purely rationally [Bendor et al., 2003], a person's view on how much influence their vote has naturally impacts their decision to vote. The individual's view is measured by multiple features in our dataset. Crepaz [1990] found that polarization of political parties lead to higher voting turnouts. Therefore, a person may be more inclined to vote when they like/identify with one party but not the other, or when they have a clear preference for one candidate. Diverse features aim to measure these inclinations. Rosenstone [1982] showed that economic adversity impacts the voting turnout. Economic problems, e.g., measured by rating of governmental economic policy or current economy, reduce a person's capacity to attend to politics and hence, participate in the elections.

Other features. We don't find any obvious causal link between voting and a specific party preference, especially within political topics. For example, it is unclear in which way preferring the Democrats on the topic of pollution and the environment impacts the decision to vote. Similarly, we don't see any direct causal link between voting and a person's political opinion on specific topics, e.g., the importance of gun control, allowing abortion, or defense spending.

A.1.4 Hypertension

The task is to diagnose whether a person has hypertension (high blood pressure) [Gardner et al., 2023]. The domains are defined by the BMI category, with people classified as overweight or obese as out-of-domain.[4] We provide a list of all features in Appendix E.3.2.

Causal features. Aging has a marked effect on the cardiovascular system and hence, increases the risk of hypertension [McEniery et al., 2007]. It is also well established that men have a higher prevalence of hypertension compared with women (prior to the onset of menopause) [Ramirez and Sullivan, 2018]. Some researchers attribute this difference to women having contact with healthcare systems more frequently [Leng et al., 2015]. Prevalence of hypertension also differs across racial/ethnic groups [Bell et al., 2010, Dorans et al., 2018].[5] We thus regard the demographic features age, sex and race as causal. Cigarette smoking is associated with an acute increase in blood pressure, mainly through stimulation of the sympathetic nervous system. Several research studies suggest that it increased the risk of hypertension [Mills et al., 2020, Virdis et al., 2010], and even cessation of chronic smoking does not lower blood pressure [Gerace et al., 1991]. Therefore, we classify former smoking as a causal feature. Moreover, patients diagnosed with diabetes are at a higher risk to also develop hypertension [Petrie et al., 2018].

Arguably causal features. The individual's current lifestyle affects their risk for hypertension, e.g., obesity, alcohol consumption, smoking, physical inactivity and unhealthy diet [Virdis et al., 2010, Mills et al., 2020]. At the same time, patients diagnosed with hypertension might cease the harmful behaviors to improve their health. Therefore, we view features encoding these behaviors as arguably causal. Researchers have long established that low socio-economic status, e.g., poverty and employment, increases the risk of hypertension [Leng et al., 2015]. When a person indeed develops hypertension, their situation may even worsen.

Anti-causal features. Some researchers regard hypertension as risk factor for the development of certain types of cancer [Sionakidis et al., 2021, Pandey et al., 2023].[6]

A.1.5 Hospital mortality

The task is to classify whether an ICU patient expires in the hospital during their current visit [Gardner et al., 2023]. The domains are defined by the insurance type, with being insures by Medicare being out-of-domain. We provide a list of all features in Appendix E.6.2.

Causal features. Walicka et al. [2021] showed that the in-hospital non-surgery-related mortality rate significantly increased with age. Averbuch et al. [2022] found evidence that sex and ethnicity are independently associated with the risk of inpatient mortality. They argue that the finding are possibly results from differences in care received in the hospital, e.g., role of bias in assessing medical risk. Therefore, we regard a person's age, sex and ethnicity as causal. Soffer et al. [2022] discuss the 'obesity paradox', i.e. medical ward patients with severe obesity have a lower risk for mortality compared to patients with normal BMI, measured by a person's height and weight upon entering the ICU.

Arguably causal features. We ask a medical practitioner working in an ICU unit for help in selecting the most important vitals that are routinely checked. While they are proxies of the patient's health, the selected vitals are also paramount in deciding the medical treatment received and therefore, the risk of in-hospital mortality.

[4] BMI measures nutritional status in adults. It is defined as a person's weight in kilograms divided by the square of the person's height in meters. Check out the WHO recommendations for more details.

[5] There is no evidence that racial and ethnic disparities in risk of hypertension are explained by genetic factors though [Kaufman et al., 2015].

[6] The causal relationship between hypertension and cancer is a prime example for major epistemic uncertainty. They are so intricately linked that they inspired their own research field [Pandey et al., 2023].

A.1.6 ASSISTments

The task is to predict whether a student solves a problem correctly on first attempt in an online learning tool [Gardner et al., 2023]. The domains are different schools. We provide a list of all features in Appendix E.5.1.

Causal features. We regard all features encoding information of the problem as causal. For example, the skill associated with the problem, the type of problem, the number of hints, and how it is framed in the online learning tool. When a student asked for a hint or solves a problem in tutor mode, the system automatically marks it as incorrect. Therefore, the target directly depends on the first action of the student, that is, whether a student asks for a hint or an explanation.

Arguably causal features. The system predicts the student's concentration, boredom, confusion and frustration. While research established that learners' cognitive-affective state influences their performance [Baker et al., 2010], the system's predictions are at best proxies of their true state of mind.

Other features. We don't see a clear link to the time in milliseconds for the student's responses.

A.2 Outlook: Classification tasks that motivate causal modeling

We highlight two prediction tasks where recent research suggests that they benefit from causal theory [Subbaswamy and Saria, 2019, Schulam and Saria, 2017].

Cooper et al. [1997] built a predictor for the probability of death for patients with pneumonia. The goal was to identify patients at low risk that can be treated safely at home for pneumonia. Their dataset contains inpatient information from 78 hospitals in 23 U.S. States. Cooper et al. [1997] assumed that hospital-treated pneumonia patients with a very low probability of death would also have a very low probability of death if treated at home. Caruana et al. [2015] pointed out that this assumption may not hold, for example, in patients with a history of asthma. Due to the existing policy across hospitals to admit asthmatic pneumonia patients to the ICU, the aggressive treatment actually lowered their mortality risk from pneumonia compared to the general population. While Caruana et al. [2015] use this observation to argue for interpretable models, Schulam and Saria [2017] take it as motivation for causal models to ensure generalization.

In another example, Nestor et al. [2019] trained predictive models on records from the MIMIC-III database between 2001 and 2002, and tested on data of subsequent years. When the underlying clinical information system changed in 2008, this caused fundamental changes in the recorded measurements and a significant drop in prediction quality of machine learning models trained on raw data. The predictive performance, however, remained surprisingly robust after aggregating the raw features into expert-defined clinical concepts.[7] If these clinical concepts reflect causal relationships, this example may be viewed as empirical support for causal modeling.

B Details on experimental procedure, algorithms and compute resources

We go into the details of the experimental procedure, including information on the confidence intervals, in Appendix B.1. We describe the machine learning methods and their hyperparameter in Appendix B.2. Experiment run details and compute resources are provided in Appendix B.3.

B.1 Experimental procedure

In this section, we go into the details of the experimental procedure and visualize the steps for three selected tasks.

First, we define up to four sets of features based on domain knowledge and common sense: all features, causal features, arguably causal features and anti-causal features. The sets of features are provided for each task in Appendix E. Exemplary explanations are provided for seven tasks. See Section 2.1 and Appendix A.

Second, we split the full dataset into in-domain set and out-of-domain set. We have a train/test/validation split within the in-domain set, and a test/validation split within the out-of-domain set. We

[7] Various measurements of the same biophysical quantity are grouped together.

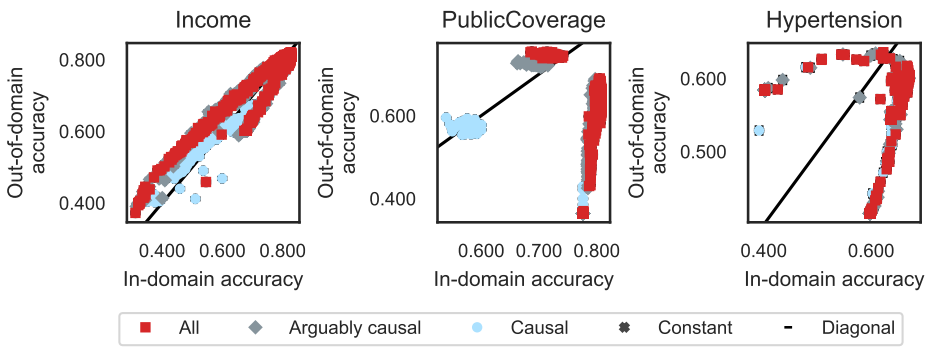


Figure 7: In-domain and out-of-domain performances of *all* trained classifier.

adapt the split choices from *TableShift*, that is 80%/10%/10% split in-domain and 90%/10% split out-of-domain.

For each feature set:

1. We apply the machine learning methods listed in Section 2.3. For each method:
   (a) We conduct a hyperparameter sweep using HyperOpt [Bergstra et al., 2013]. We exclusively train on the training set, and use the in-domain validation accuracy for hyperparameter tuning. A method is tuned for 50 trials. Note that we withhold the out-domain validation set from training, i.e., domain adaption is not possible.
   (b) We evaluate the trained classifiers using accuracy on in-domain and out-of-domain test set. In total, we train $3 \cdot 12 \cdot 50 = 1,800$ classifier for tasks with one training domain, and $3 \cdot 23 \cdot 50 = 3,450$ classifiers for tasks with at least two training domains. See Figure 7.
   (c) We select the best model according to their in-domain validation accuracy. This follows the selection procedure in Gulrajani and Lopez-Paz [2020] and Gardner et al. [2023]. See Figure 8. To ensure compatibility, we add the best in-domain and out-of-domain accuracy pair observed by Gardner et al. [2023]. See Figure 9. We restrict our further analysis to this selection.
2. We find the Pareto-set $\mathcal{P}$ of in-domain and out-of-domain accuracy pairs. See Figure 10. We dismiss Pareto dominant classifiers whose in-domain accuracy is smaller than the constant prediction, i.e., predict worse than the majority prediction in-domain. This is the final selection of classifiers. We also add the Pareto-dominated region. See Figure 11. We compute the shift gaps, and find the Pareto-set of shift gap and out-of-domain accuracy of the set $\mathcal{P}$.

Note that we do not use the out-of-domain validation set in our experiments. It is left for further analysis, e.g., measuring the shift between in-domain and out-of-domain [Gardner et al., 2023].

Error bounds. We use 95% Clopper-Pearson confidence intervals for accuracy. They attain the nominal coverage level in an exact sense. We approximate 95% confidence intervals for the shift gap, the difference between in-domain and out-of-domain accuracy. The Clopper-Pearson confidence intervals for the accuracy $\hat{p}$ approximately equal the normal version for large sample sizes n,

$$[l_{\text{CP}}, u_{\text{CP}}] \approx [\hat{p} - z_{0.975}\sigma, \hat{p} + z_{0.975}\sigma]$$

with z_α the α-quantile of the standard normal distribution and $\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ the standard error.

We can immediately infer confidence intervals for the shift gap, as the in-domain and out-of-domain test sets are independent. The variance of the shift gap is the sum of the variance of in-domain and

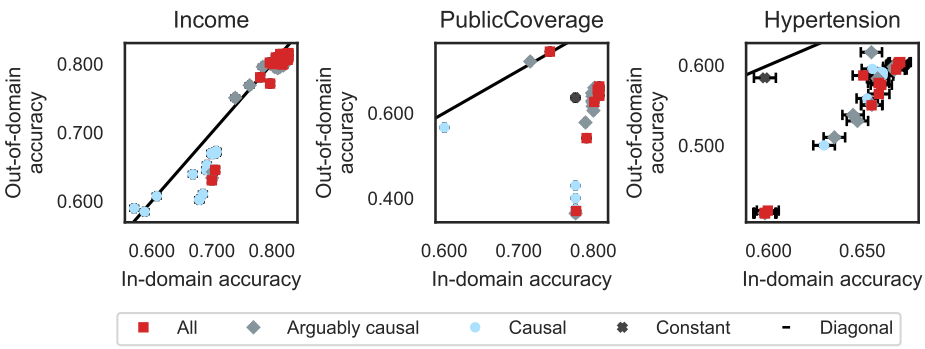


Figure 8: In-domain and out-of-domain performances of trained classifier with best in-validation accuracy *within* a model class.

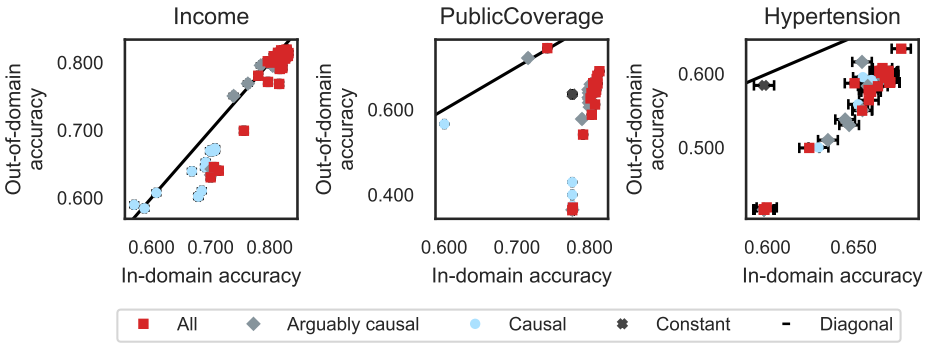


Figure 9: In-domain and out-of-domain performances of trained classifier with best in-validation accuracy *within* a model class, as well as results obtained by Gardner et al. [2023]

out-of-domain accuracy. We therefore compute approximate confidence intervals as

$$\left[\hat{\Delta} - \sqrt{(l_{\text{CP,in-domain}} - \hat{p}_{\text{in-domain}})^2 + (l_{\text{CP,out-of-domain}} - \hat{p}_{\text{out-of-domain}})^2},\right.$$
$$\left.\hat{\Delta} + \sqrt{(u_{\text{CP,in-domain}} - \hat{p}_{\text{in-domain}})^2 + (u_{\text{CP,out-of-domain}} - \hat{p}_{\text{out-of-domain}})^2}\right]. \tag{7}$$

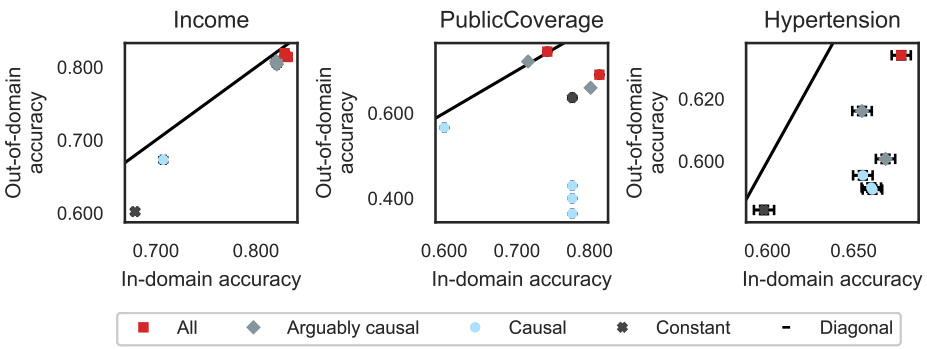


Figure 10: In-domain and out-of-domain performances of Pareto dominant classifier.

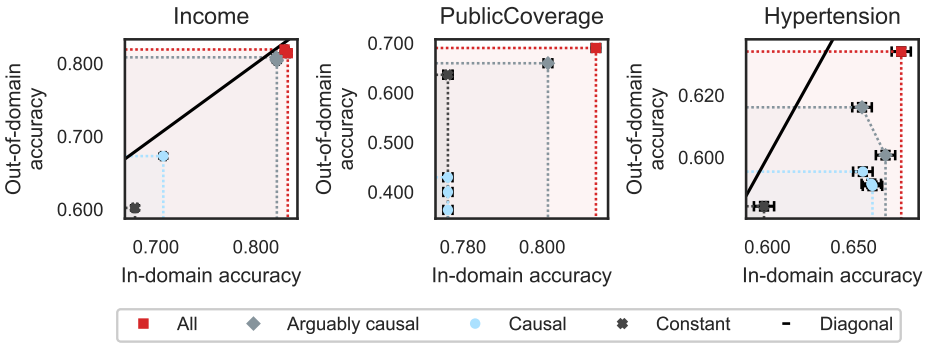


Figure 11: In-domain and out-of-domain performances of the final selection of classifiers, i.e. Pareto dominant classifier whose in-domain accuracy is better than the constant prediction. Dotted lines indicate Pareto frontiers and shaded areas the Pareto-dominated sets.

B.2 Machine learning algorithms and hyperparameters

We describe the causal methods, and refer the reader to Gardner et al. [2023] for the other machine learning algorithms. All the causal methods are motivated by causal theory and seek to find some form of invariance across multiple training domains.

Invariant Risk Minimization (IRM). IRM [Ahuja et al., 2022] modifies the training objective to learn feature representation such that the optimal linear classifier that maps the representation to the target is the same across domains.

Risk Extrapolation (REx). REx [Krueger et al., 2021] seeks to reduce variances in risk across training domains, in order to gain robustness to distributional shifts.

Information Bottleneck IRM (IB-IRM). IB-IRM [Ahuja et al., 2022] augments IRM with an information bottleneck constraint. The constraint resolves some issues of IRM.

Causal Invariant Representation Learning (Causal IRL). Causal IRL [Chevalley et al., 2022] proposes a regularizer that enforces invariance through distribution matching. We train both versions of the algorithm, i.e. with CORAL and MMD.

ANDMask. ANDMask [Parascandolo et al., 2021] is an algorithm based on the logical AND. It aims to focus on invariances and prevents memorization.

Table 2: Hyperparameter grids of causal methods

Model	Hyperparameter	Values
IRM	IRM λ	LogUniform$(1e-1, 1e5)$
	IRM Penalty Anneal Iters	LogUniform$(1, 1e4)$
REx	REx λ	LogUniform$(1e-1, 1e5)$
	REx Penalty Anneal Iters	LogUniform$(1, 1e4)$
IB-IRM	IRM λ	LogUniform$(1e-1, 1e5)$
	IRM Penalty Anneal Iters	LogUniform$(1, 1e4)$
	IB λ	LogUniform$(1e-1, 1e5)$
	IB Penalty Anneal Iters	LogUniform$(1, 1e4)$
CausIRL	MMD γ	Uniform$(1e, 1e1)$
ANDMask	ANDMask τ	Uniform$(0.5, 1)$

Table 3: Summary of trained and evaluated models. We train 23 models for tasks with at least 2 training domains, and 12 models for tasks with one training domain. Main results include the models trained on causal, arguably causal and all features.

	#Task	#Models	#Trials	Total
Main results	3×8	23	50	27,600
	3×8	12	50	14,400
Anti-causal	2	23	50	2,300
	3	12	50	1,800
Causal discovery	4	23	50	4,600
	1	12	50	600
Misclassification tests	100	23	50	115,000
	92	12	50	55,200
Random subset tests	14×500	4	10	280,000
Ablation study	234	4	10	9,360
Total				510,860

The hyperparameters are chosen with HyperOpt [Bergstra et al., 2013]. We provide the hyperparameter grid for the causal methods in Table 2. The hyperparameter grids are adapted from Gardner et al. [2023] and Gulrajani and Lopez-Paz [2020].

B.3 Experiment run details

All experiments were run as jobs submitted to a centralized cluster, running the open-source HTCondor scheduler. Each job was given the same computing resources: 1 CPU. Compute nodes use AMD EPYC 7662 64-core CPUs. Memory was allocated as required for each task: all jobs were allocated at least 128GB of RAM; for the tasks 'Public Coverage' jobs were allocated 384GB of RAM. An experiment job accounts for training and evaluating a single model for a given tasks and feature selection. Jobs were terminated when the runtime exceeded 4 hours.

We typically train and evaluate 600 models per task with a single training domain and 1,150 models per task with at least two training domains, for each feature selection. A total of 42K models were trained for the main results, and 468K models for additional results and robustness tests. We detail the number of trained models in Table 3. Preliminary experiments merely required a negligible amount of compute, and were run on a local computer.

We use the implementation of HyperOpt [Bergstra et al., 2013] in the *TableShift* API to sample from the hyperparameter space of the model. Detailed descriptions and hyperparameter choices are found in Appendix B.2 and Gardner et al. [2023].

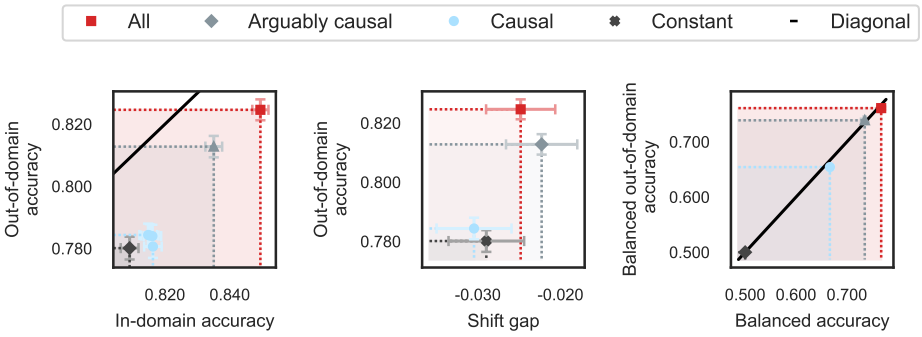
We provide the complete code base to replicate our experiments under `https://github.com/socialfoundations/causal-features`.

C Details on empirical results

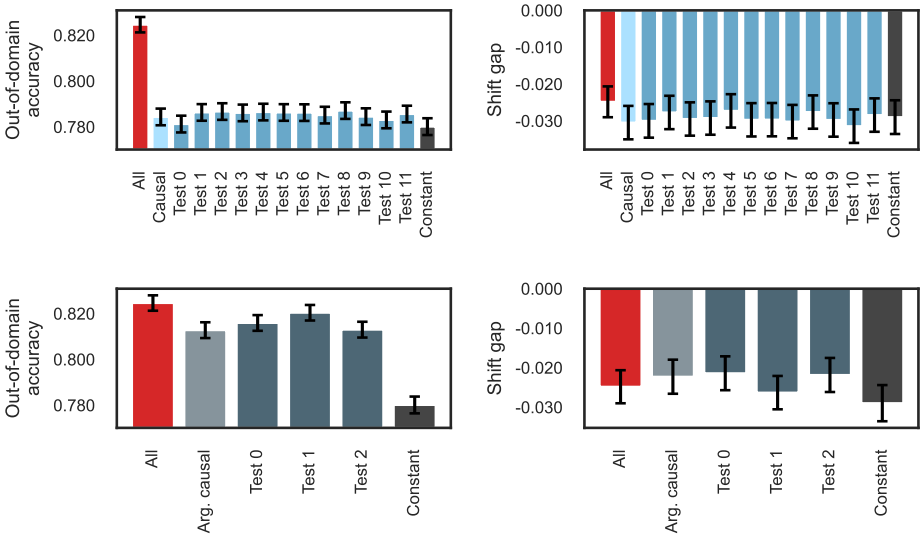
In this section, we provide figures and details of our empirical results. First, we provide the main empirical results in Appendix C.1, and results on anti-causal features in Appendix C.2. Then, we describe and show results from the causal machine learning algorithms and causal discovery methods in Appendix C.3 and Appendix C.4, respectively. We give details and results on the robustness test involving random subsets in Appendix C.5. We follow with an ablation study in Appendix C.6. Last, we add details on the different machine learning models to the main results in Appendix C.7.

C.1 Main empirical results

We show main results for all tasks in Figures 12 - 27. More precisely, we provide: (1) the Pareto-frontiers of in-domain and out-of-domain accuracy by feature selection; (2) the Pareto-frontiers of shift gap and out-of-domain accuracy by feature selection; (3) Pareto-frontiers of in-domain and out-of-domain *balanced* accuracy by feature selection; and (4) out-of-domain accuracies and shift gaps obtained by robustness tests for causal and arguably causal features.

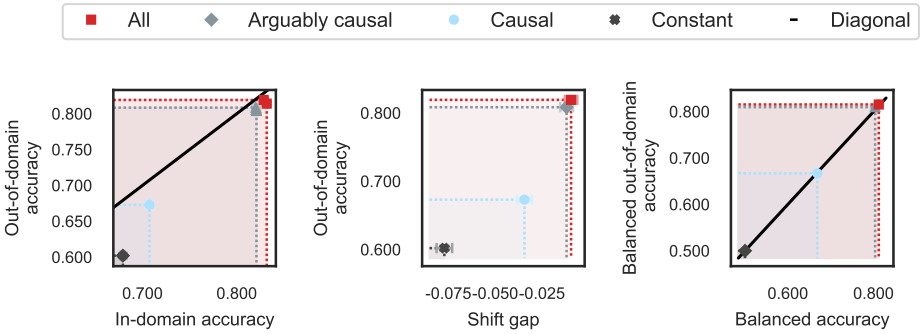


(a) Pareto-frontiers by feature selection.

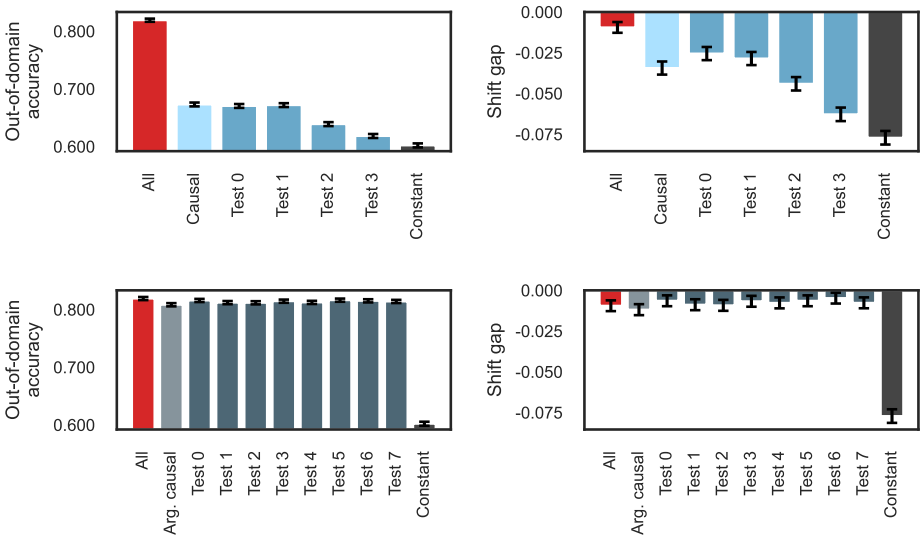


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 12: Food Stamps

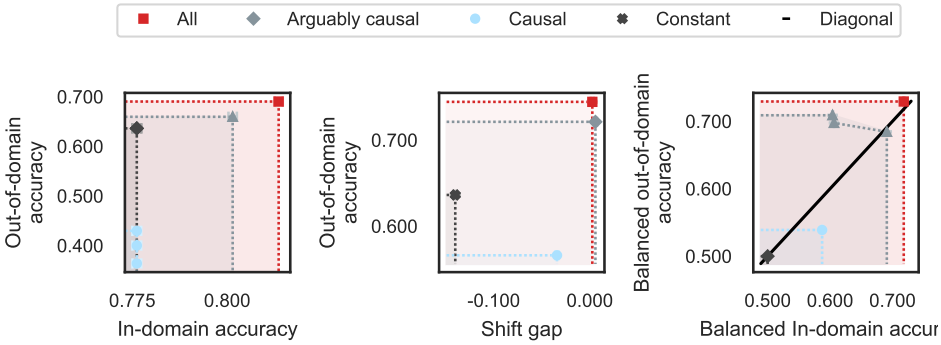


(a) Pareto-frontiers by feature selection.

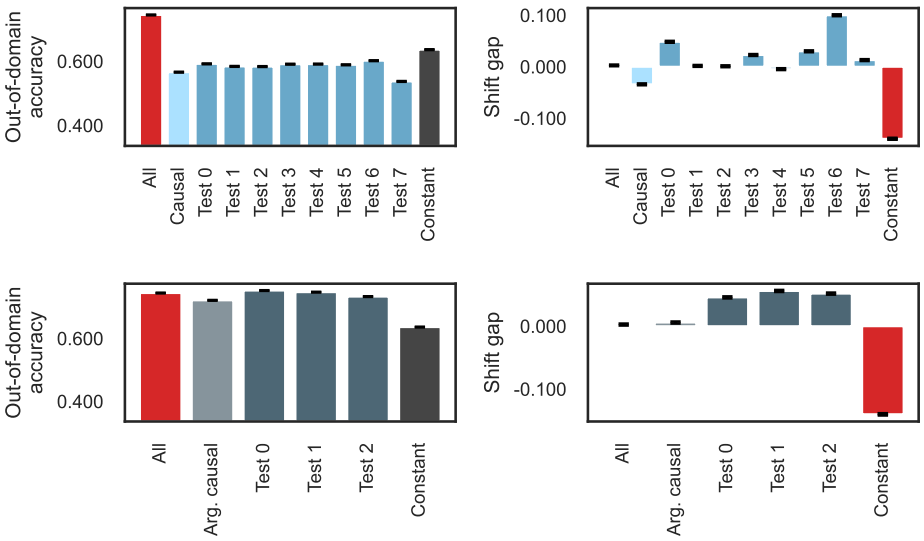


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 13: Income

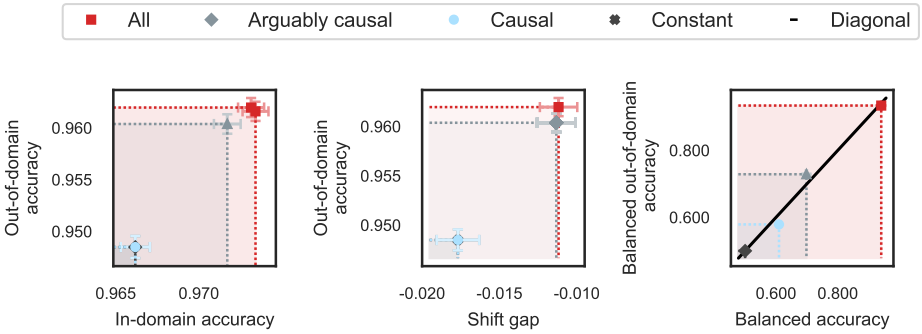


(a) Pareto-frontiers by feature selection.

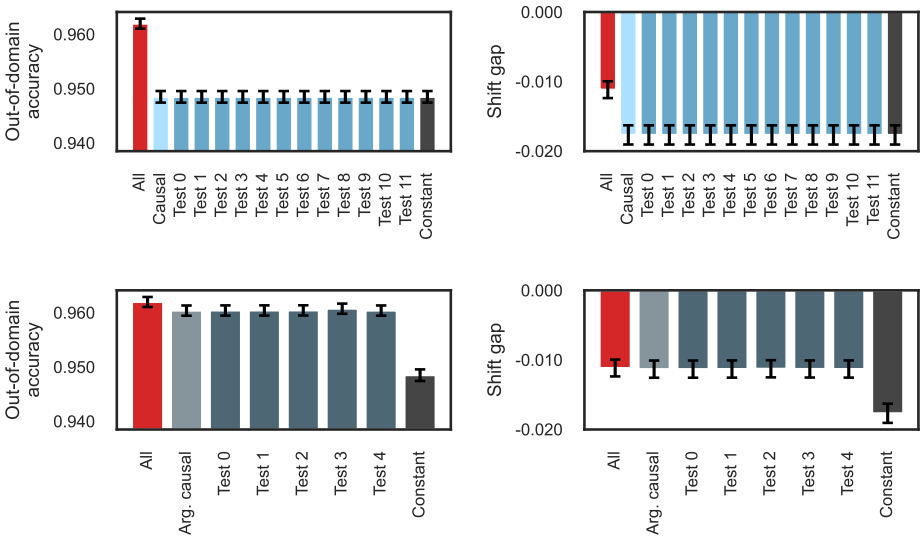


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 14: Public Coverage

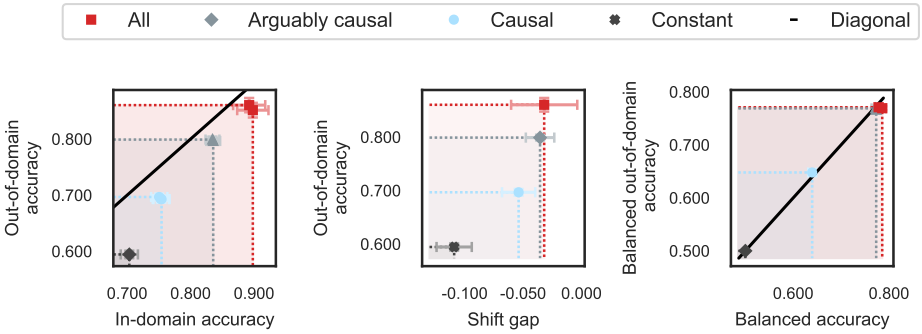


(a) Pareto-frontiers by feature selection.

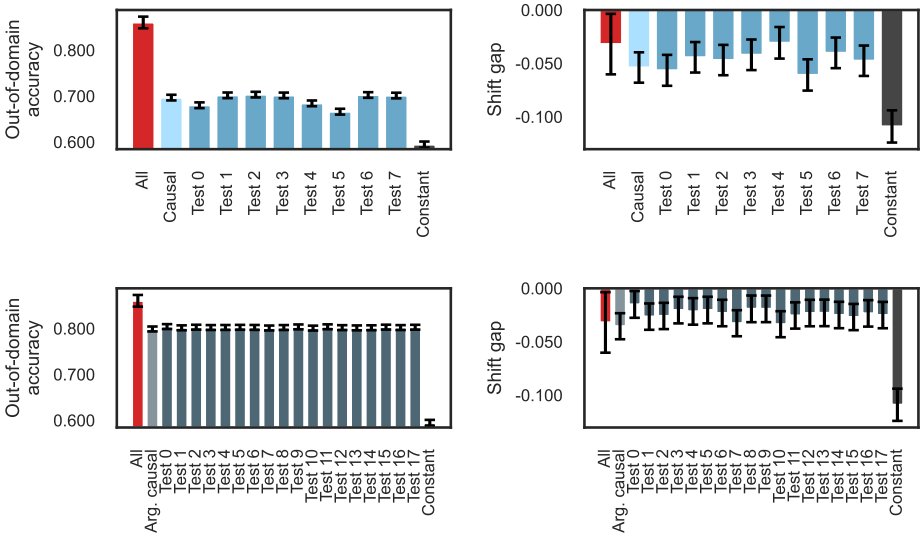


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 15: Unemployment

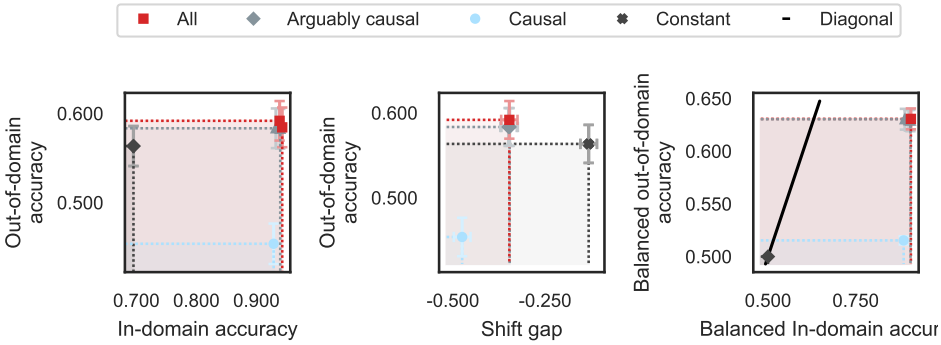


(a) Pareto-frontiers by feature selection.

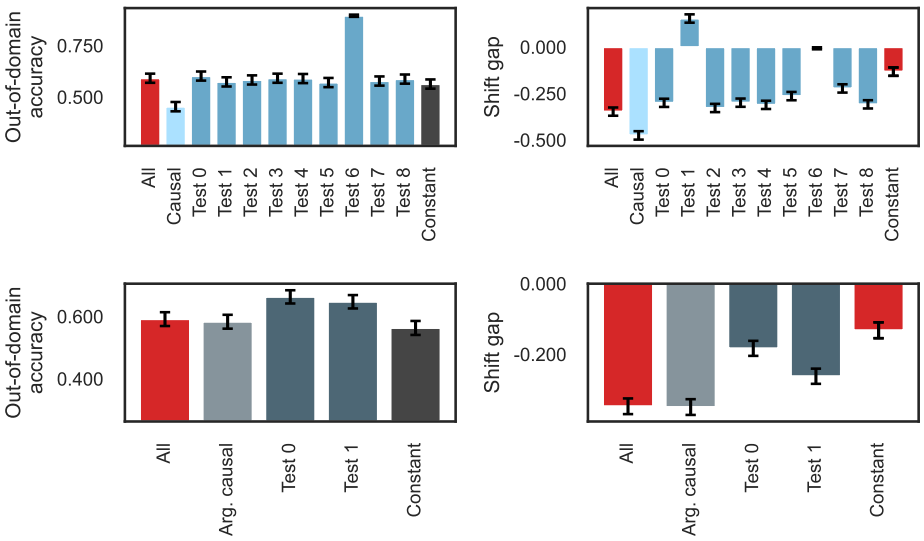


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 16: Voting

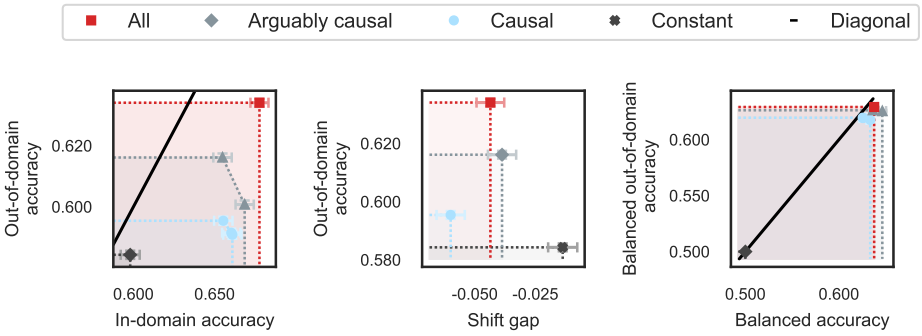


(a) Pareto-frontiers by feature selection.

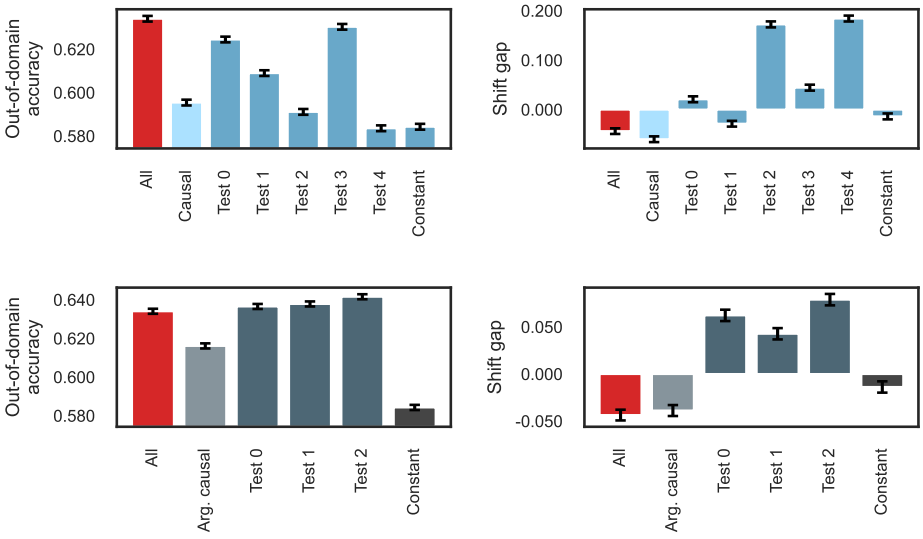


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 17: ASSISTments

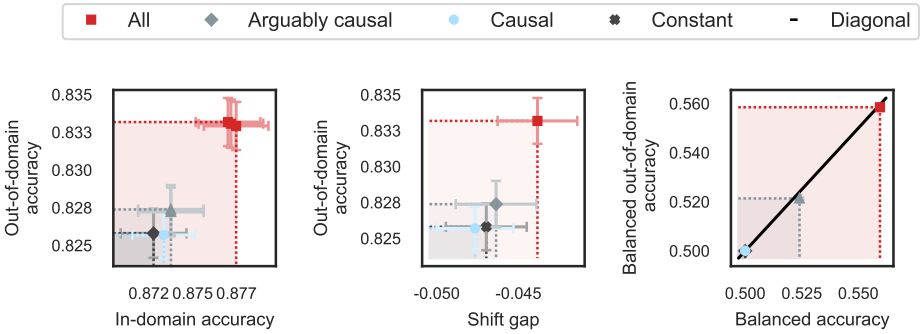


(a) Pareto-frontiers by feature selection.

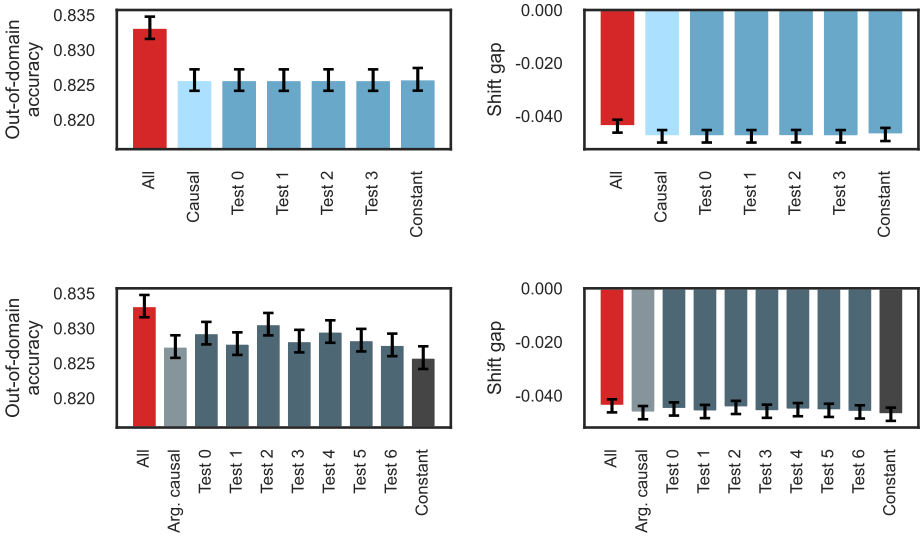


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 18: Hypertension

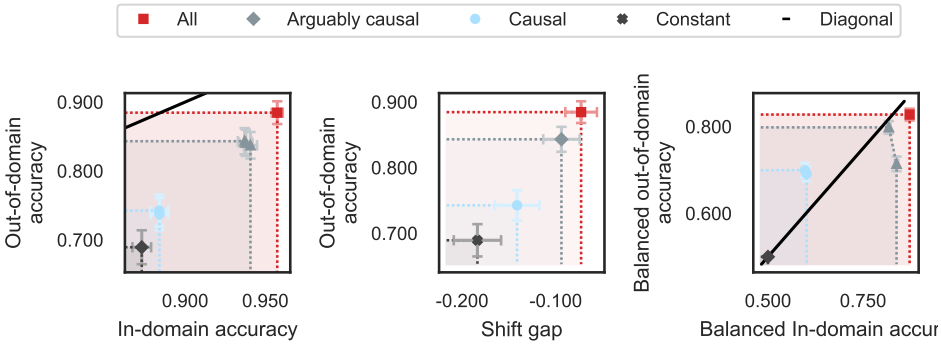


(a) Pareto-frontiers by feature selection.

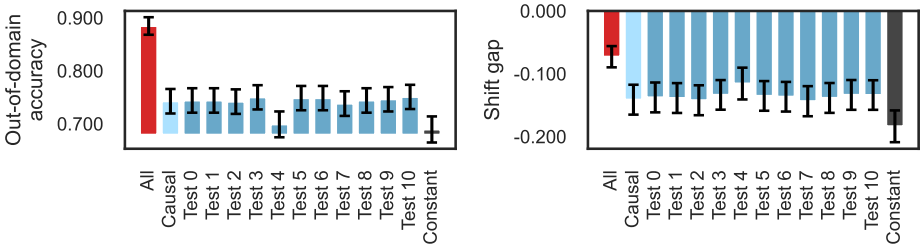


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 19: Diabetes

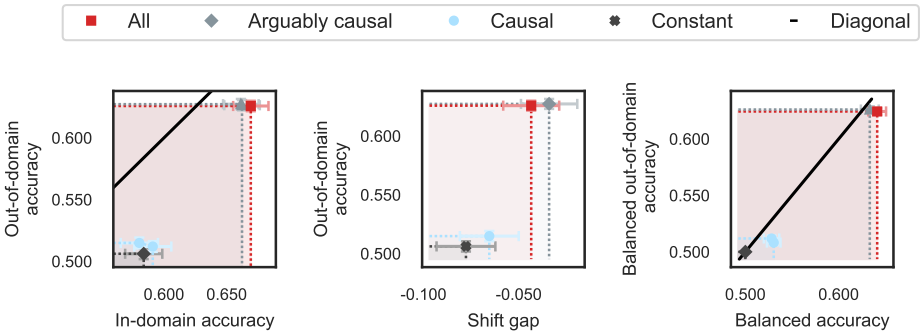


(a) Pareto-frontiers by feature selection.

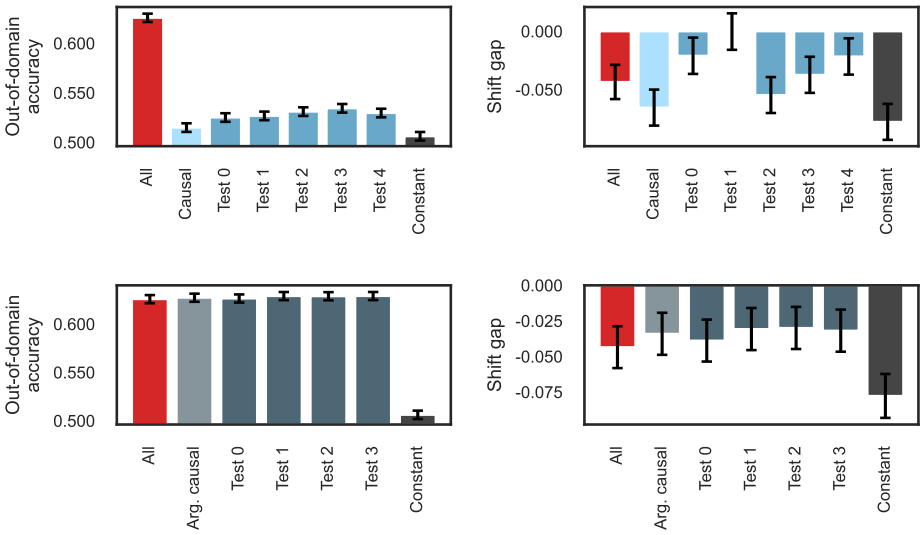


(b) Robustness tests for causal features.

Figure 20: College Scorecard

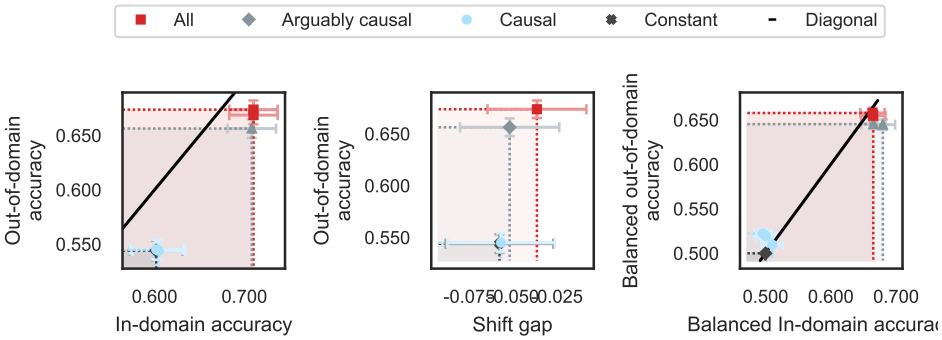


(a) Pareto-frontiers by feature selection.

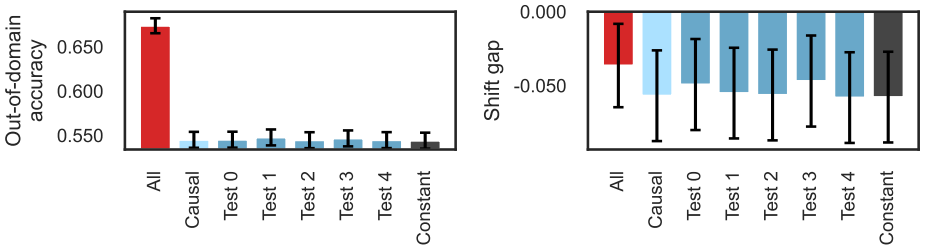


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 21: Hospital Readmission

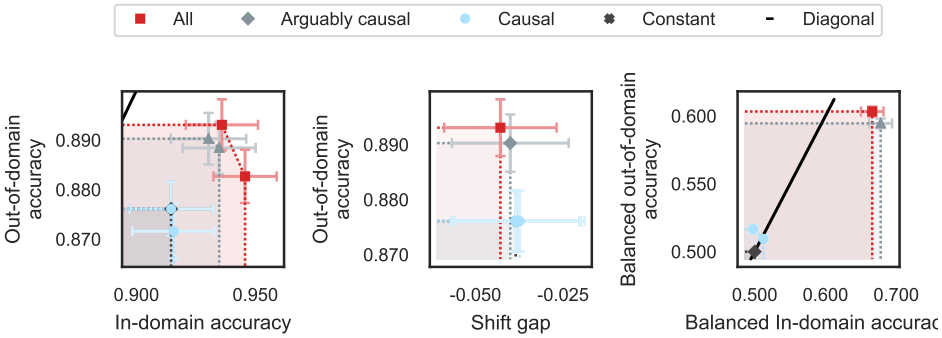


(a) Pareto-frontiers by feature selection.

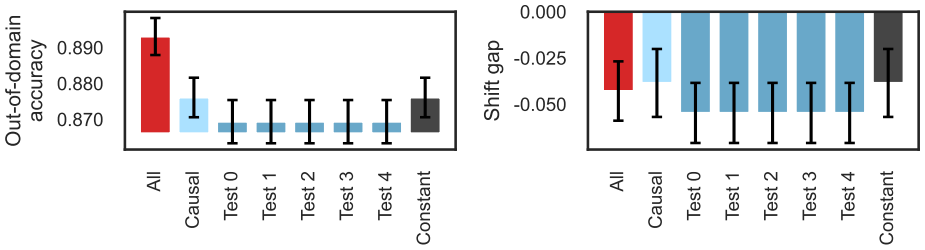


(b) Robustness tests for causal features.

Figure 22: Stay in ICU

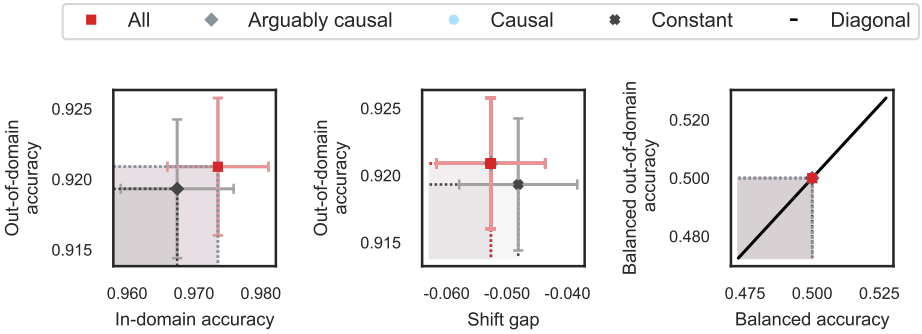


(a) Pareto-frontiers by feature selection.

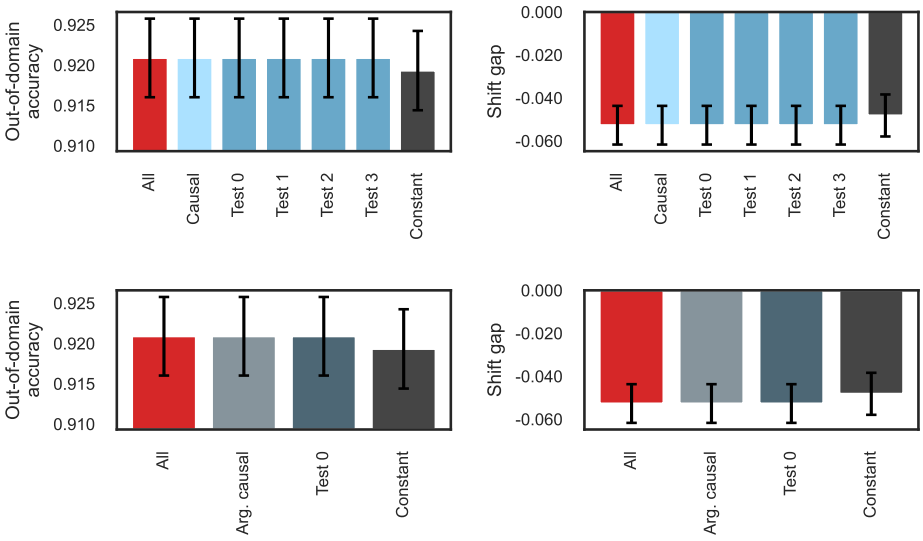


(b) Robustness tests for causal features.

Figure 23: Hospital Mortality

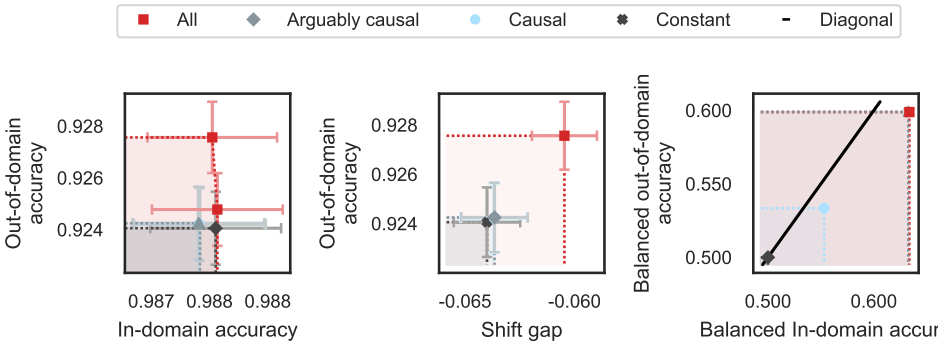


(a) Pareto-frontiers by feature selection.

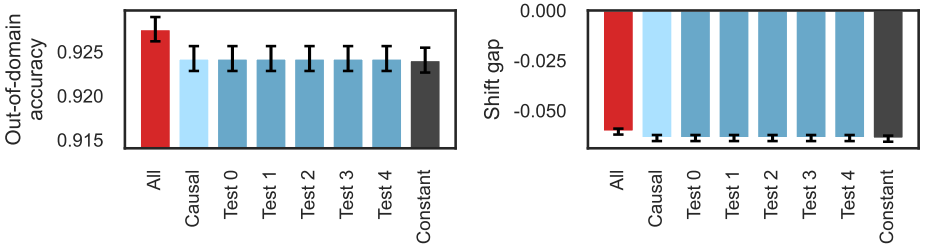


(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 24: Childhood lead

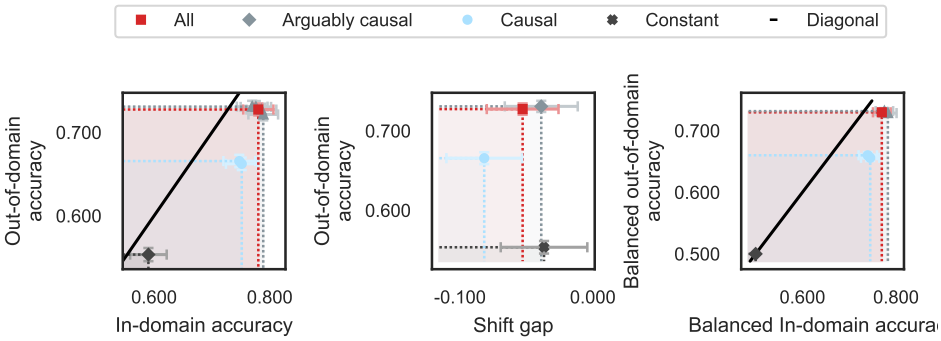


(a) Pareto-frontiers by feature selection.

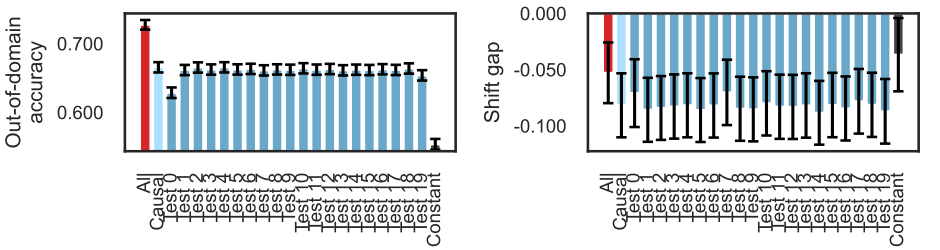


(b) Robustness tests for causal features. No robustness test for arguably causal features as there is only one additional features.

Figure 25: Sepsis

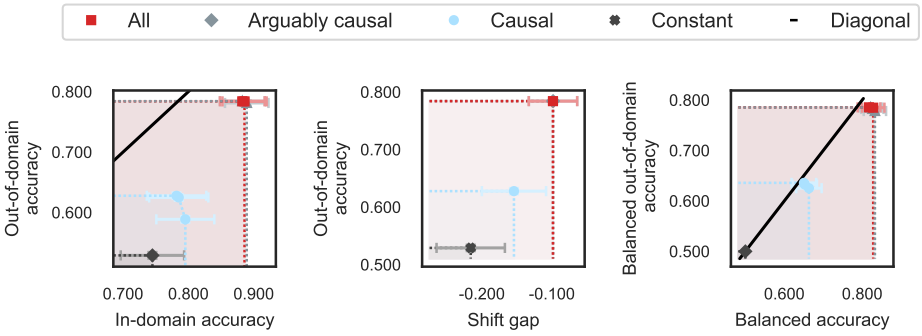


(a) Pareto-frontiers by feature selection.

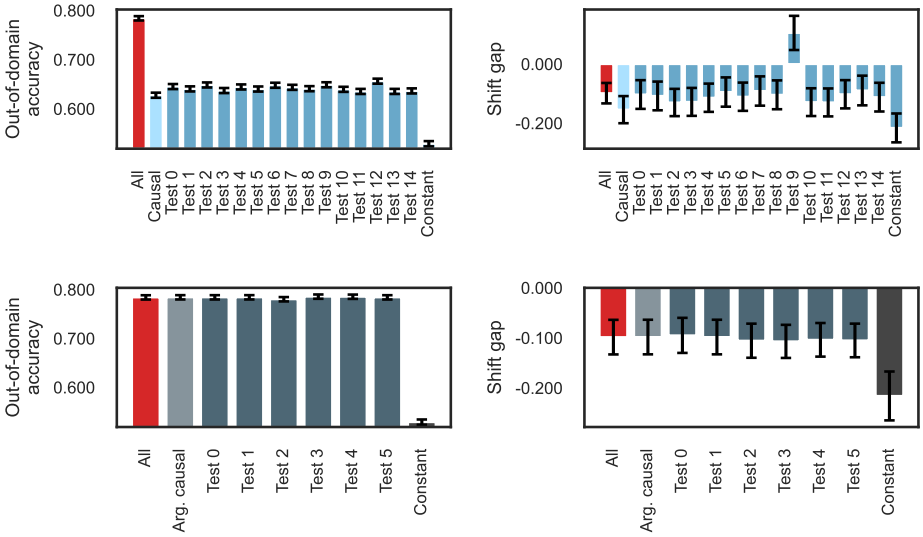


(b) Robustness tests for causal features.

Figure 26: Utilization



(a) Pareto-frontiers by feature selection.



(b) Robustness tests for causal features (upper) and arguably causal features (lower).

Figure 27: Poverty

C.2 Anti-causal features

We have five tasks in which some features are plausibly anti-causal: 'Income', 'Unemployment', 'Diabetes', 'Hypertension' and 'Poverty'. Figures 28 and 29 show the Pareto frontiers of anti-causal feature in comparison to the causal selections and all features. We also show the Pareto frontiers we achieve by training on arguably causal and anti-causal features.

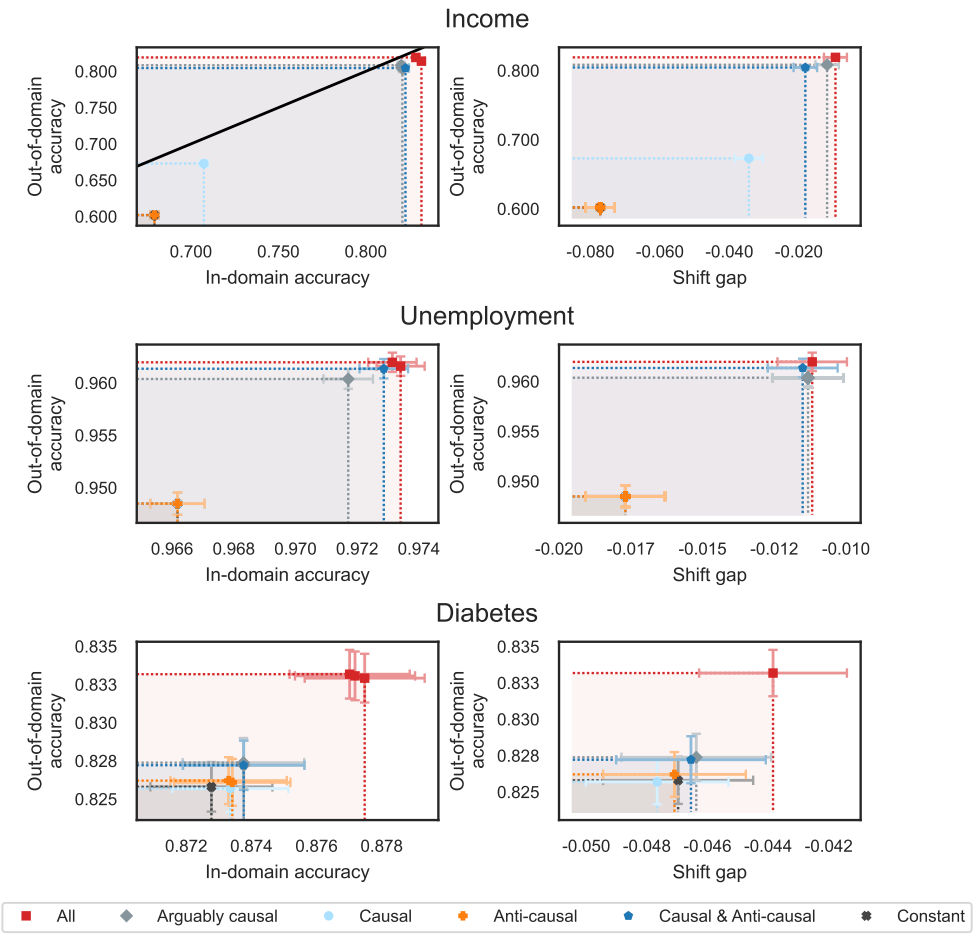


Figure 28: (Left) Pareto-frontiers of in-domain and out-of-domain accuracy of anti-causal features in comparison to causal features sets and all features. (Right) Pareto-frontiers of shift gap and out-of-domain accuracy attained.

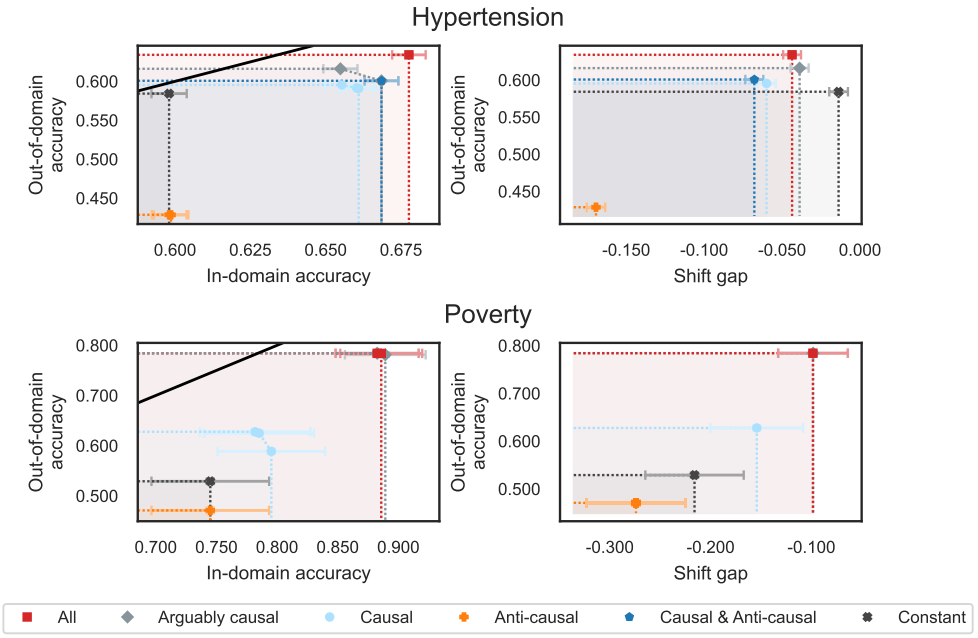


Figure 29: (Left) Pareto-frontiers of in-domain and out-of-domain accuracy of anti-causal features in comparison to causal features sets and all features. (Right) Pareto-frontiers of shift gap and out-of-domain accuracy attained. (Continued)

C.3 Causal machine learning methods

We evaluate five causal methods: Invariant Risk Minimization (IRM) [Arjovsky et al., 2019], Risk Extrapolation (REx) Krueger et al. [2021], Information Bottelneck IRM (IB-IRM) [Ahuja et al., 2022], Causal IRL based on CORAL and MMD [Chevalley et al., 2022] and AND-Mask [Parascandolo et al., 2021]. A description and the hyperparameter grids are given in Appendix B.

The causal methods require at least two testing domains with each a sufficient amount of data. Eight of our task satisfy these requirements: 'Food Stamps', 'Income', 'Unemployment', 'Voting', 'College Scorecard', 'Hospital Readmission', 'Hospital Mortality' and 'Length of Stay'.
Note that the task 'ASSISTments' is not included. It has multiple training domains but very few data point in some of them.

We provide results in Figure 31 and 32. The bar plot in Figure 30 summarized how often the performance is: (i) smaller than the performance of the causal features, (ii) similar to the performance of causal features, (iii) between the performances of the causal features and arguably causal features, and (iv) similar to the performance of the arguably causal features. Note that the causal machine learning algorithm never outperform the arguably causal features.

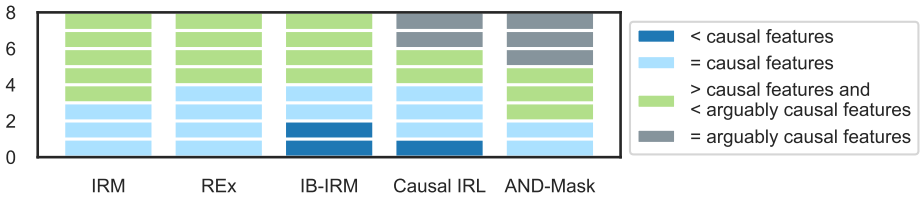


Figure 30: Performance of causal methods in comparison to causal and arguably causal features. Summary for the 8 tasks with multiple training domains.

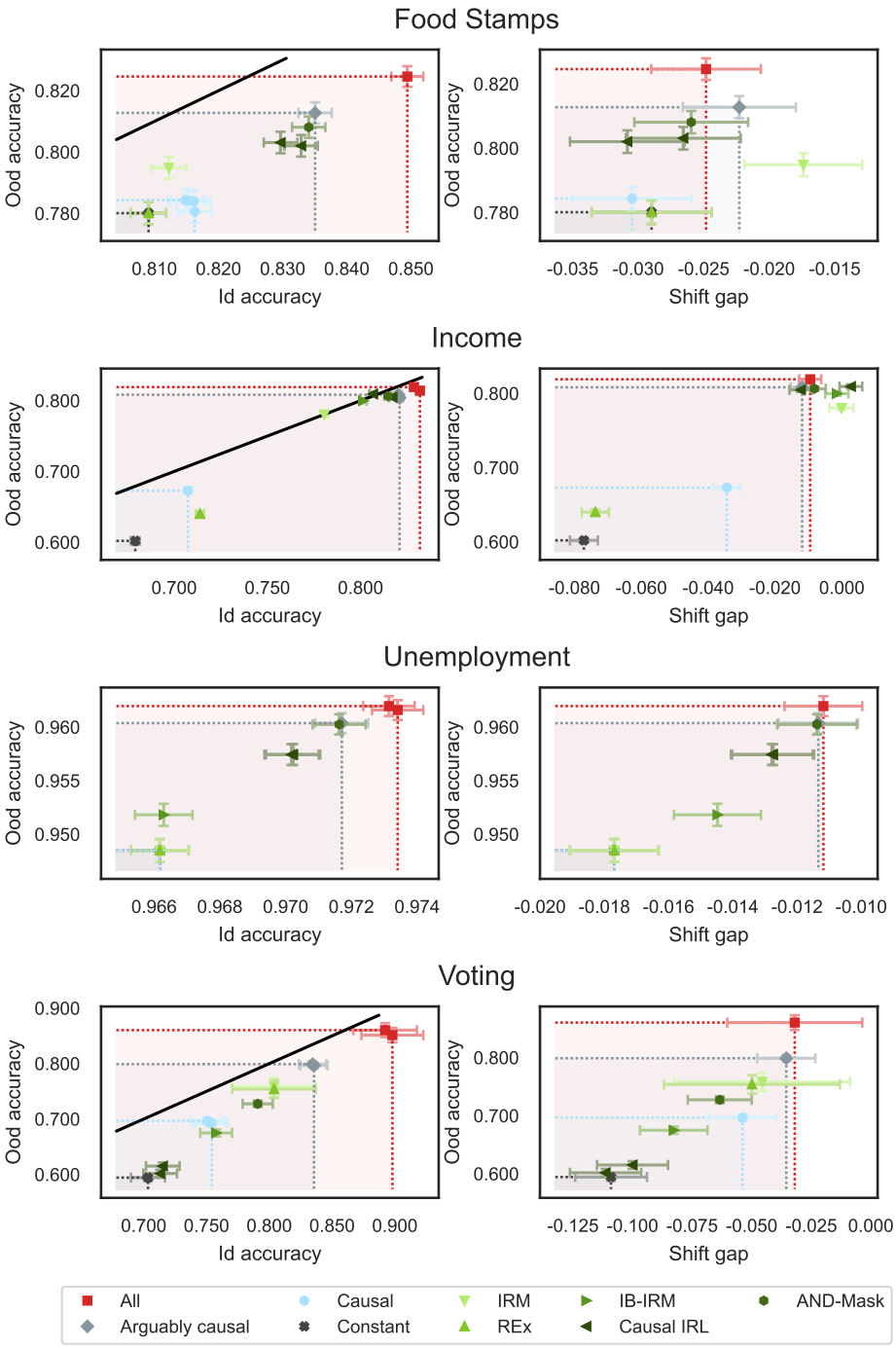


Figure 31: (Left) Pareto-frontiers of in-domain and out-of-domain accuracy of causal methods and domain-knowledge features selection. (Right) Pareto-frontiers of shift gap and out-of-domain accuracy attained.

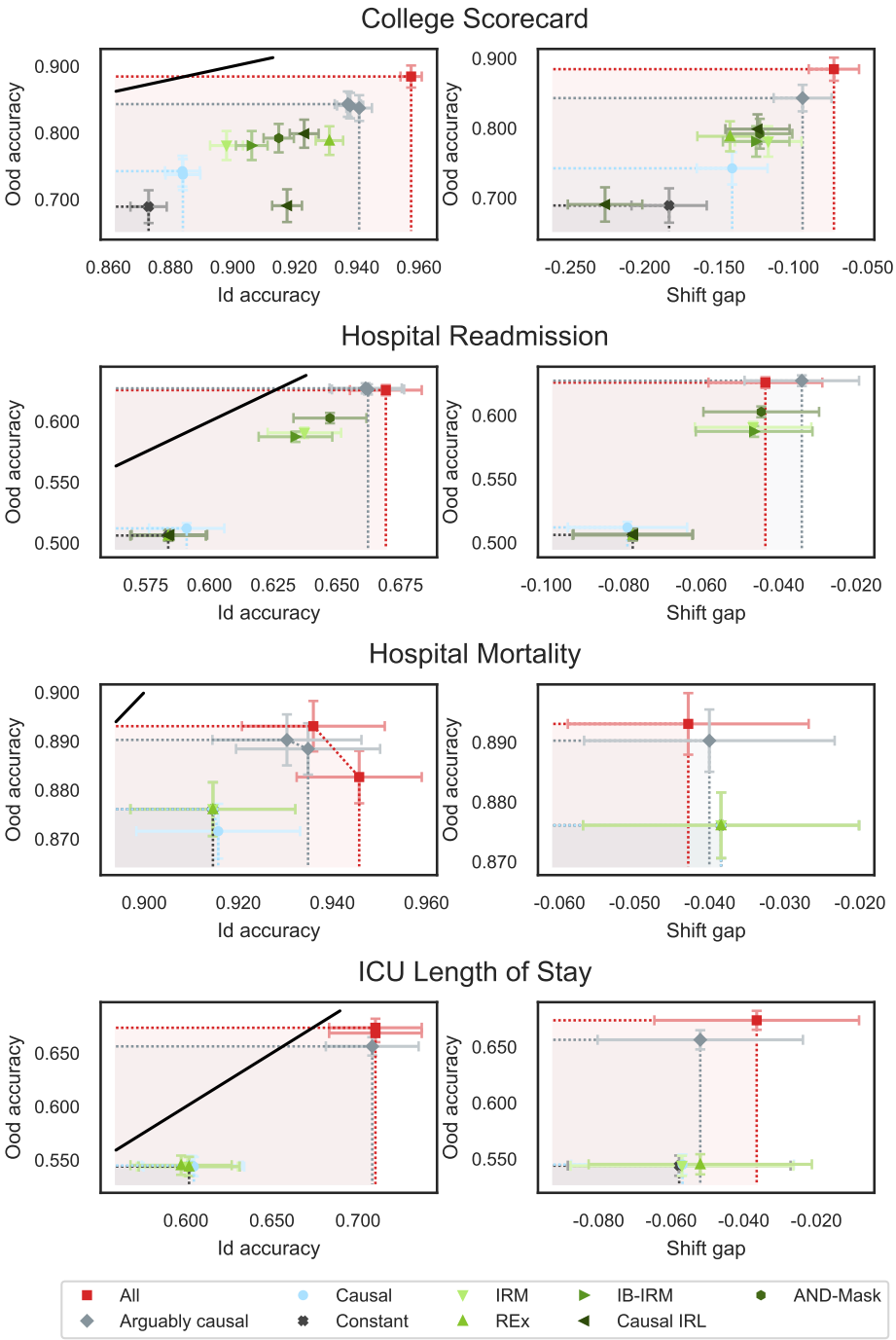


Figure 32: (Left) Pareto-frontiers of in-domain and out-of-domain accuracy of causal methods and domain-knowledge features selection. (Right) Pareto-frontiers of shift gap and out-of-domain accuracy attained. (Continued)

C.4 Causal discovery algorithms

We consider invariant causal prediction (ICP) [Peters et al., 2016] and classic causal discovery methods. In addition, we contemplated score matching methods. When benchmarked to other causal discovery methods, they showed surprising robustness in settings where assumptions on the data may be violated [Montagna et al., 2024]. We first describe results of causal discovery methods we analyzed in our experiments. Then, we explain the challenges we encountered with the score matching methods and the reason we didn't include them in our final analysis.

We list the evaluated causal discovery methods and describe their results.

Invariant Causal Prediction (ICP). ICP [Peters et al., 2016] collects all subset of features that show invariance in their predictive accuracy across domains, and outputs valid confidence intervals for the causal relationships. The variables with an effect significantly different from zero are the causal predictors under sufficient assumptions. ICP requires at least two training domains, each needs a sufficient amount of data. In addition, the number of features needs to be of a reasonable size. This is provided in 6 tasks: 'Food Stamps', 'Income', 'Unemployment', 'Voting', 'College Scorecard' and 'Hospital Readmission'. We use the boosting implementation from the R package 'InvariantCausalPrediction' by Peters et al. [2016]. We choose a confidence level of $\alpha = 0.05$; it is the default setting.

Peter-Clark algorithm (PC). The PC algorithm [Spirtes et al., 2000] is a classical causal discovery method. It is based on conditional independence testing and estimates a completed partially directed acyclic graph (CPDAG). We use the implementation from the R package 'pcalg' by Kalisch et al. [2012] with the default confidence level of $\alpha = 0.01$. We do not consider the tasks 'Hospital Mortality' and 'Stay in ICU' due to computational costs.

Fast causal inference algorithm (FCI). The FCI algorithm [Spirtes et al., 1995] is a generalization of the PC algorithm. It allows arbitrarily many latent and selection variables. It outputs a partial ancestral graph (PAG). We use the implementation from the R package 'pcalg' by Kalisch et al. [2012] with a confidence level of $\alpha = 0.01$. We do not consider the tasks 'Hospital Mortality' and 'Stay in ICU' due to computational costs.

Standard conditional independence tests assume one common data type, that is, either binary, discrete or Gaussian. Our datasets are however a mix of binary, categorical and continuous variables. We decide to bin the continuous variables into five categories, and then use a discrete independence test in the PC and FCI algorithm. Note that ICP is applied to standard preprocessed data, that is, binary, one-hot encoded categorical and continuous data.

We run the causal discovery algorithms on the in-domain validation set. If the algorithm outputs any causal parents, we train the machine learning methods listed in Section 2.3 on the training set. We tuned each method for 50 trials. We provide the results in Table 4 and 5. Estimated causal parents are denoted in square brackets. A task is *rejected* from ICP in Table 4 if no subset of variables leads to invariance across the domains. We showcase an example of a CPADAG from the PC algorithm in Figure 33. An example of a PAG from the FCI algorithm is given in Figure 34.

We provide the performance of all estimated causal parents in Figure 35 and 36.

Score matching methods and compute memory. We considered the implementation of the score matching methods provided by Montagna et al. [2024]. See `https://github.com/py-why/dodiscover`. The algorithms compute a matrix of size [sample size x sample size x features]. This is computationally infeasible when using the whole in-domain validation set. For example, this requires 2,88 TiB of memory for 'Income' and 3.09 TiB of memory for 'Unemployment'. Note that the validation sample sizes are merely 121,154 and 158,015, respectively.
A solution is to randomly sample, say, 1,000 data points from the validation set.[8] We perform preliminary experiments to assess this approach. We sample 1,000 data points, and run the SCORE [Rolland et al., 2022] and Discovery At Scale (DAS) algorithm by [Montagna et al., 2023] for the main tasks 'Diabetes', 'Income' and 'Unemployment'.

[8]This is the largest number of sample size in the analysis of Montagna et al. [2024].

Table 4: Summary of empirical results for invariant causal prediction (ICP) with $\alpha = 0.05$. The descriptions of features are given in Appendix E.

Task	#Features	Has ≥ 2 training domains with sufficient sample size	ICP
Food Stamps	28	✓	rejected
Income	23	✓	rejected
Public Coverage	19	✗	not applicable
Unemployment	16	✓	[RELP,WRK]
Voting	54	✓	no causal predictors
Diabetes	25	✗	not applicable
Hypertension	18	✗	not applicable
College Scorecard	118	✓	rejected
ASSISTments	15	✗	not applicable
Stay in ICU	7491	✓	not applicable
Hospital Mortality	7491	✓	not applicable
Hospital Readmission	46	✓	rejected
Childhood Lead	7	✗	not applicable
Sepsis	40	✗	not applicable
Utilization	218	✗	not applicable
Poverty	54	✗	not applicable

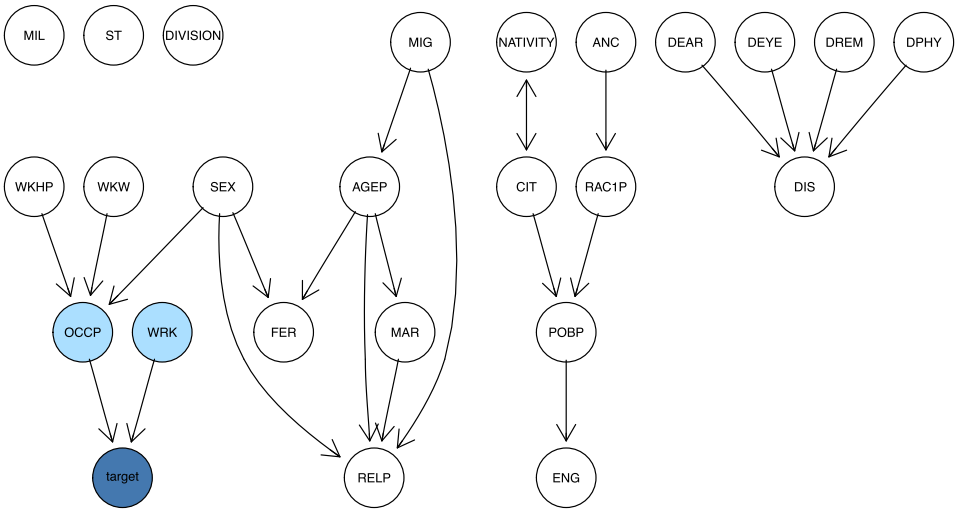


Figure 33: CPDAG estimated by the PC algorithm for the task 'Unemployment'. The target denotes the employment states. The descriptions of features are given in Appendix E.

The computation of the SCORE algorithm fails for all tasks. One computation step during the pruning does not converge after a few pruning steps.

We obtain results from the DAS algorithm. We provide the estimated DAG from DAS under the filenames das_diabetes.svg, das_income.svg and das_unemployment.svg at `https://github.com/socialfoundations/causal-features/tree/add-ons/experiments_causal/add_on_results/causal_discovery/`.

The algorithm doesn't estimate any causal parents for being diagnosed with diabetes ('Diabetes') or having a certain income level ('Income'). The only causal parent output for being unemployed is being born in South Dakota ('Unemployment'). While the results are intriguing, they hardly promise supreme prediction outcomes. Therefore, we didn't pursue the score matching methods further. We however encourage future research on checking their performance on all tasks. Another path for future research is to find more sophisticated solutions to reduce the required memory amount.

Table 5: Summary of empirical results for the PC and FCI algorithm [Spirtes et al., 1995, 2000] with $\alpha = 0.01$. The descriptions of features are given in Appendix E.

Task	#Features	PC	FCI
Food Stamps	28	[HUPAC,PUBCOV]	no causal parents
Income	23	[WKHP,AGEP,HINS1,OCCP]	no causal parents
Public Coverage	19	no causal parents	no causal parents
Unemployment	26	[OCCP,WRK]	no causal parents
Voting	54	no causal parents	no causal parents
Diabetes	25	[HIGH_BLOOD_PRESS]	[HIGH_BLOOD_PRESS]
Hypertension	18	no causal parents	no causal parents
College Scorecard	118	no causal parents	no causal parents
ASSISTments	15	no causal parents	no causal parents
Stay in ICU	7491	not applicable	not applicable
Hospital Mortality	7491	not applicable	not applicable
Hospital Readmission	46	no causal parents	no causal parents
Childhood Lead	7	no causal parents	no causal parents
Sepsis	40	no causal parents	no causal parents
Utilization	218	no causal parents	no causal parents
Poverty	54	no causal parents	no causal parents

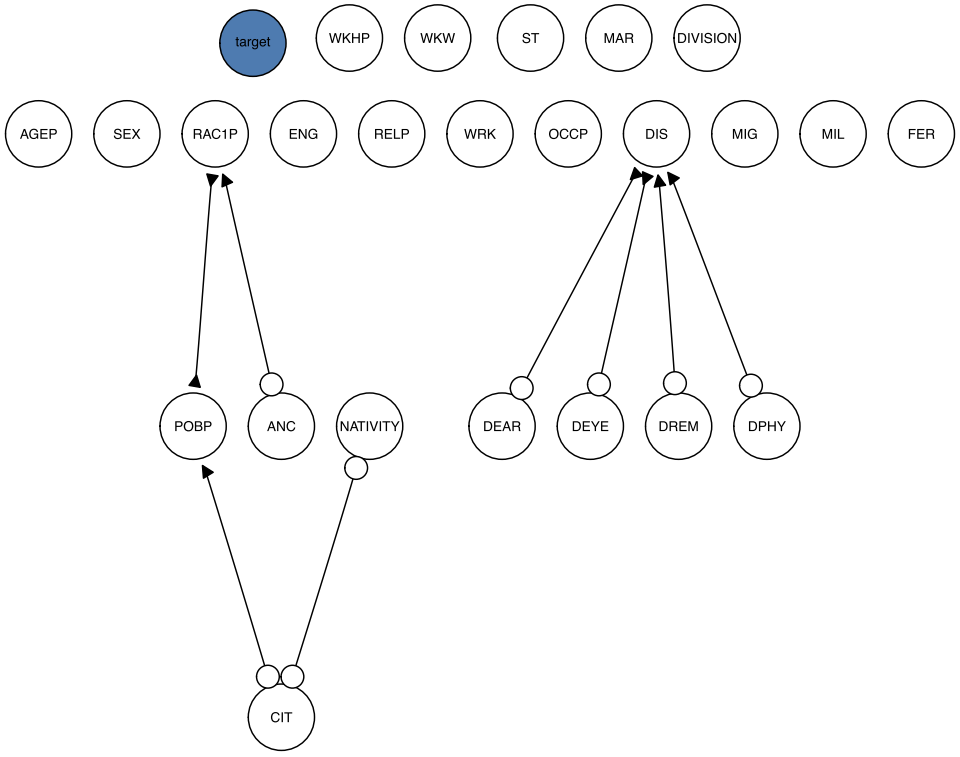
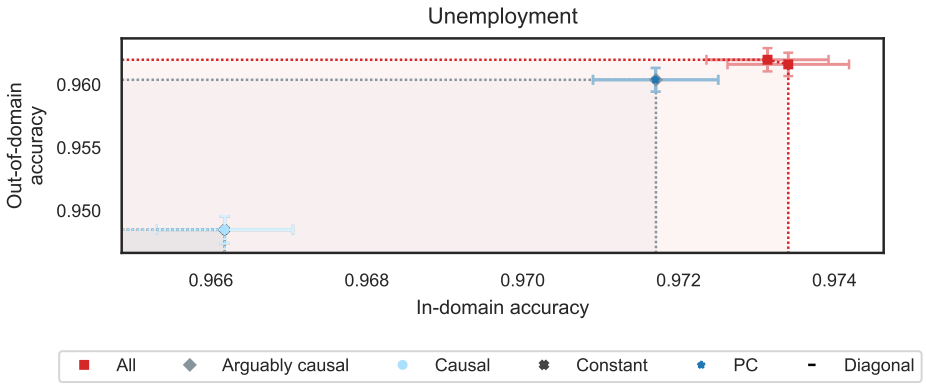
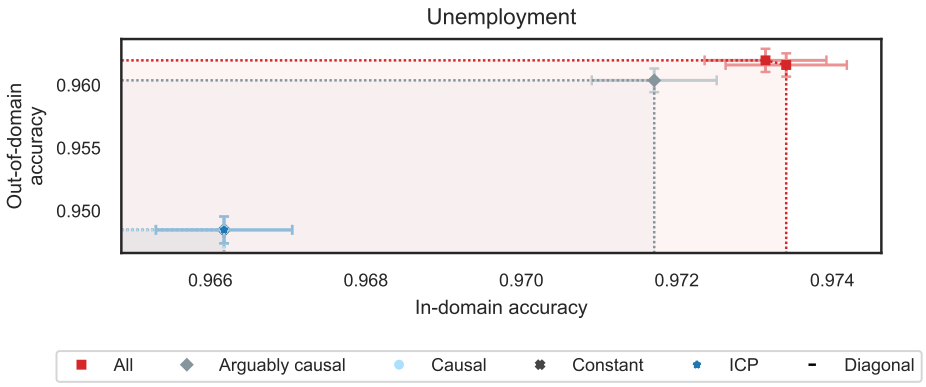


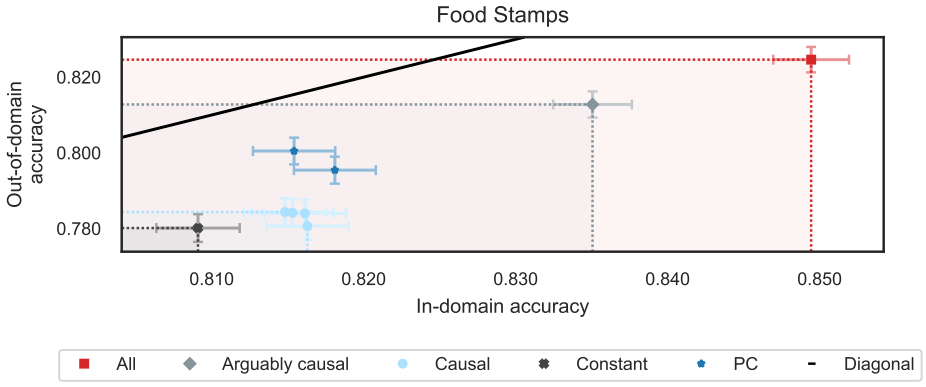
Figure 34: PAG estimated by the FCI algorithm for the task 'Unemployment'. The target denotes the employment states. The descriptions of features are given in Appendix E.



(a) PC algorithm on the task 'Unemployment'.

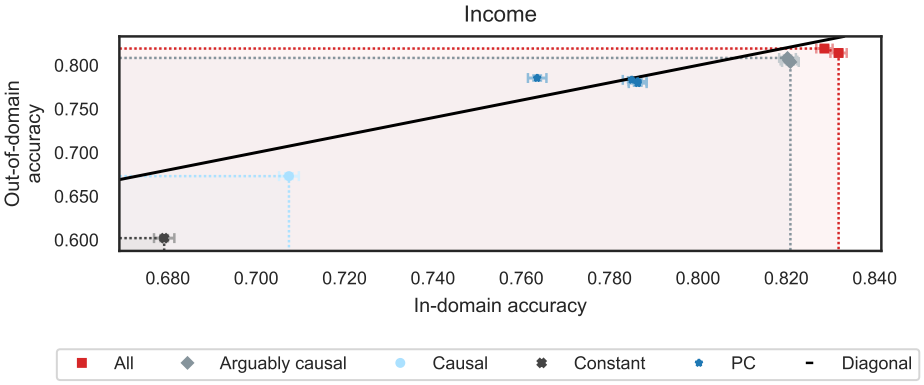


(b) ICP algorithm on the task 'Unemployment'.

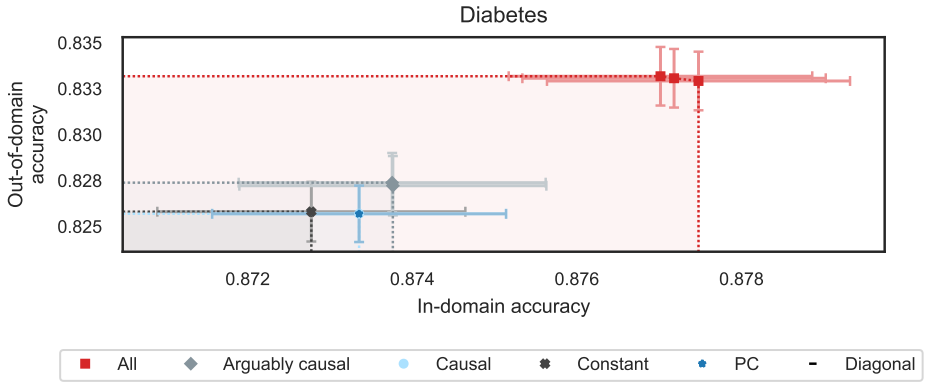


(c) PC algorithm on the task 'Food Stamps'.

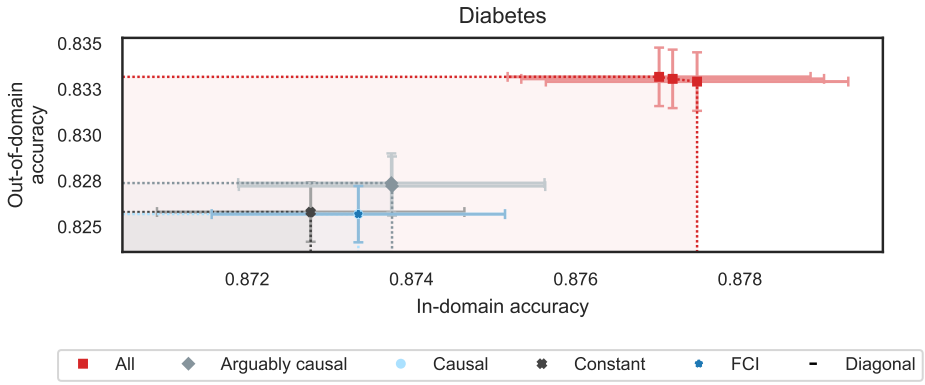
Figure 35: Performance of causal parents selected by causal discovery algorithm, in comparison to domain knowledge selected causal features and whole feature set.



(a) PC algorithm on the task 'Income'.



(b) PC algorithm on the task 'Diabetes'.



(c) FCI algorithm on the task 'Diabetes'.

Figure 36: Performance of causal parents selected by causal discovery algorithm, in comparison to domain knowledge selected causal features and whole feature set. (Continued)

C.5 Random subsets

We test whether there exists a subset of features that achieve significantly higher out-of-domain accuracy than the whole feature set. It is however computational infeasible to evaluate all possible subsets of the features for our tasks. For example, the task 'Income' with 23 features has already ≈ 8 million subsets.

We randomly sample 500 subsets for each task, with exception to 'Hospital Mortality' and 'Stay in ICU'. We don't think that 500 random sample are informative for 'Hospital Mortality' and 'Stay in ICU', as it is just a teeny fraction of the power set of features (2^{7491} subsets).

Due to computational cost, we further restrict our analysis to the models XGBoost, LightGBM, FT Transformer and SAINT. These models achieve the highest average out-of-domain accuracy across tasks. See Appendix C.7 and Gardner et al. [2023]. The methods are also tuned for 10 trials, instead of 50.

We provide the results in Figure 37, 38, 39 and 40. Except for some subsets in the task 'ASSISTments', none of our random subsets outperforms the full feature set, not in in-domain accuracy nor in out-of-domain accuracy. In the task 'ASSISTments', we predict whether a question is correct answered by a student in an online learning tool. The exception occurs when removing feature 'skill_id', encoding the type of skill required. The distribution of the feature 'skill_id' shifts significantly across schools, that is, from training schools to out-of-domain testing schools.

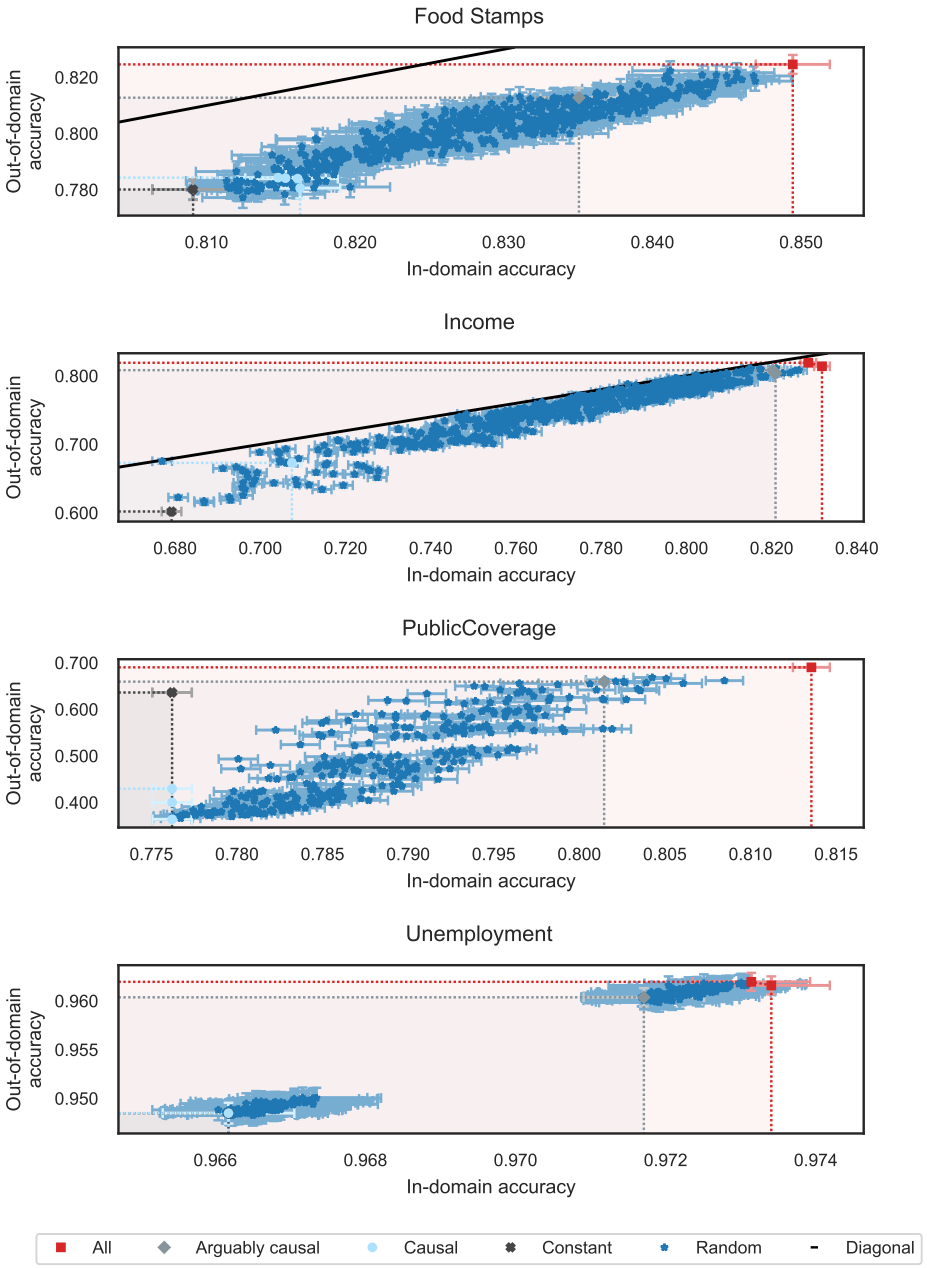


Figure 37: Performance of random subsets in comparison to causal feature selection and the whole feature set.

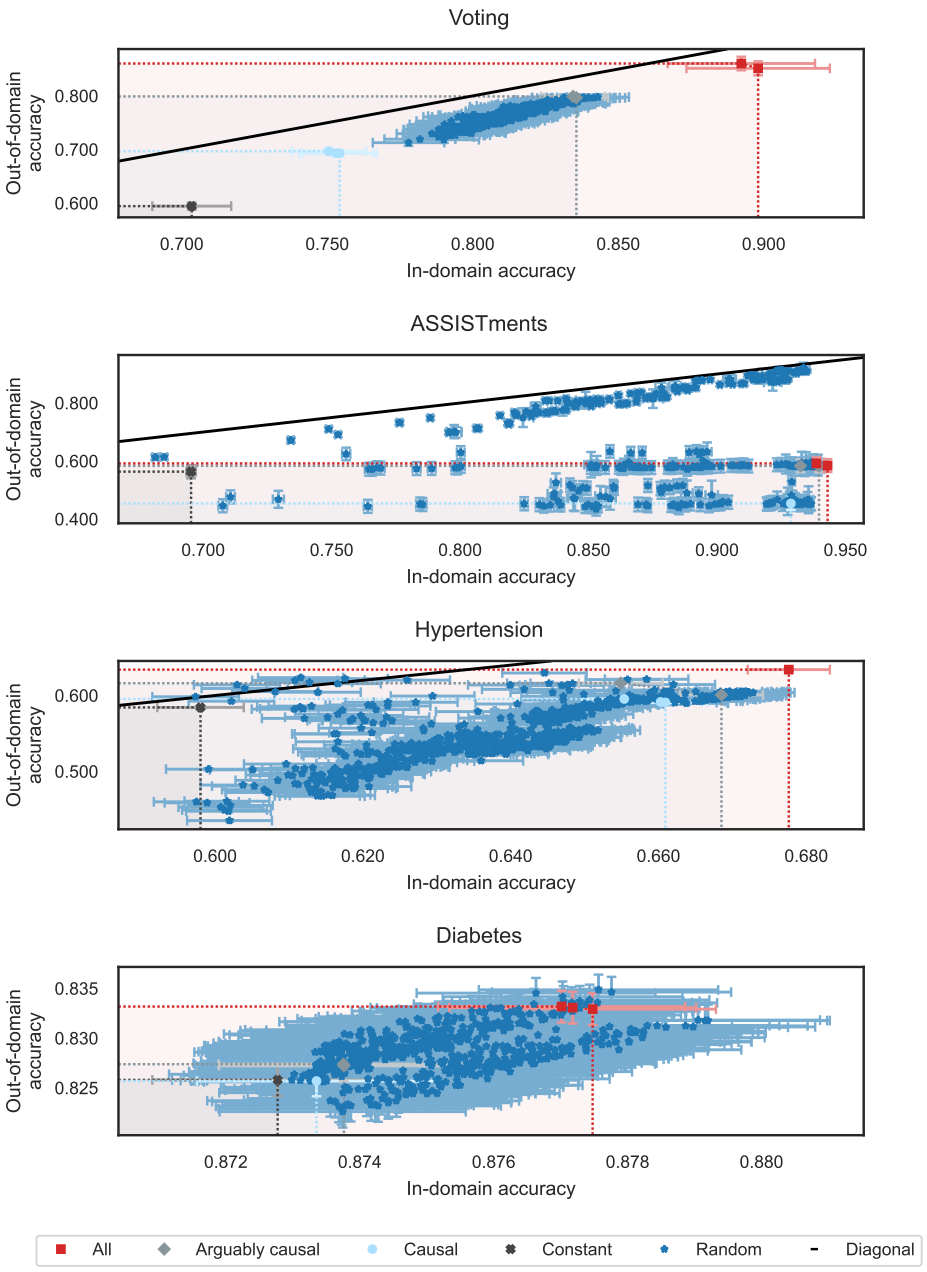


Figure 38: Performance of random subsets in comparison to causal feature selection and the whole feature set. (Continued)

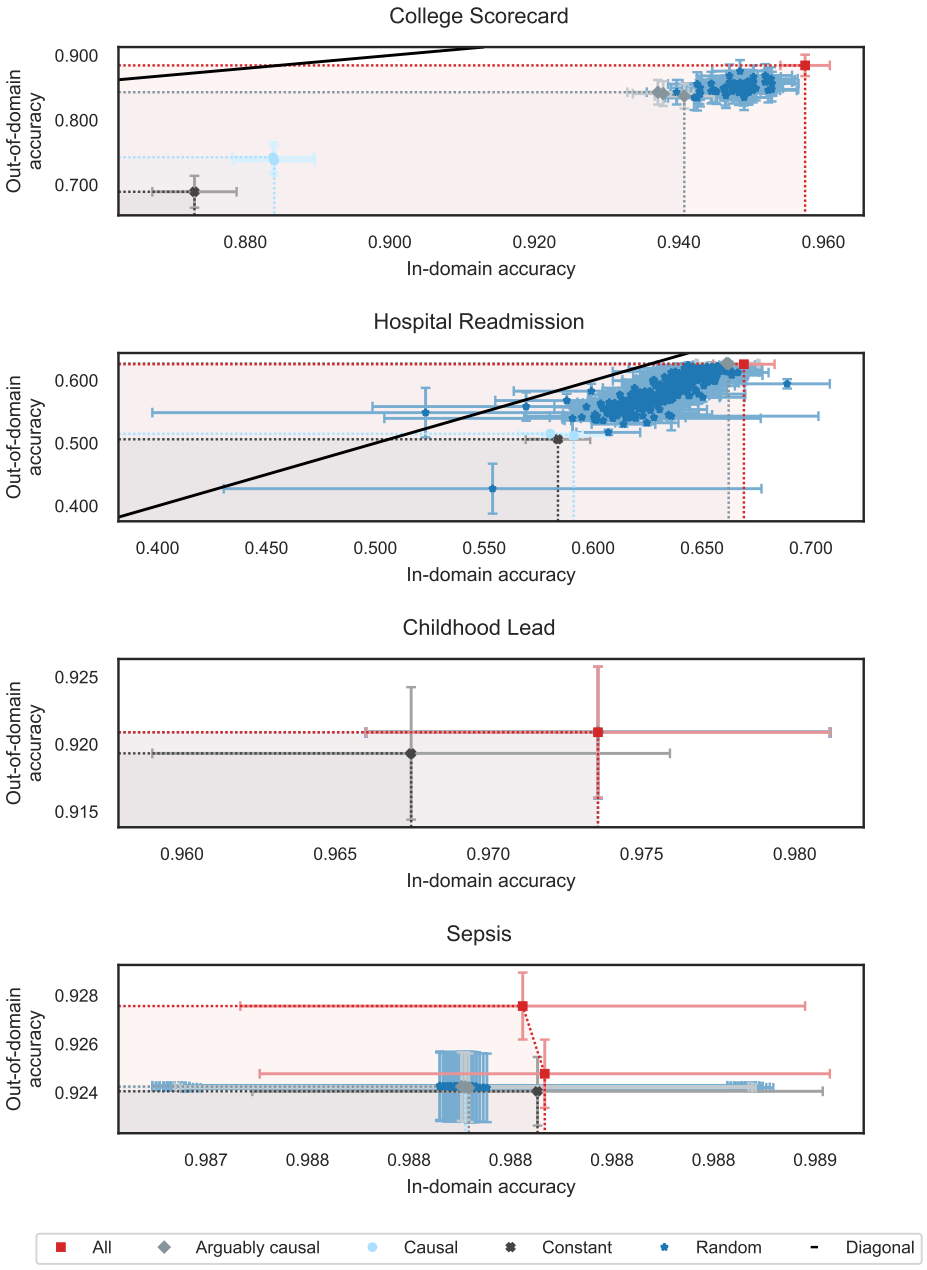


Figure 39: Performance of random subsets in comparison to causal feature selection and the whole feature set. (Continued)

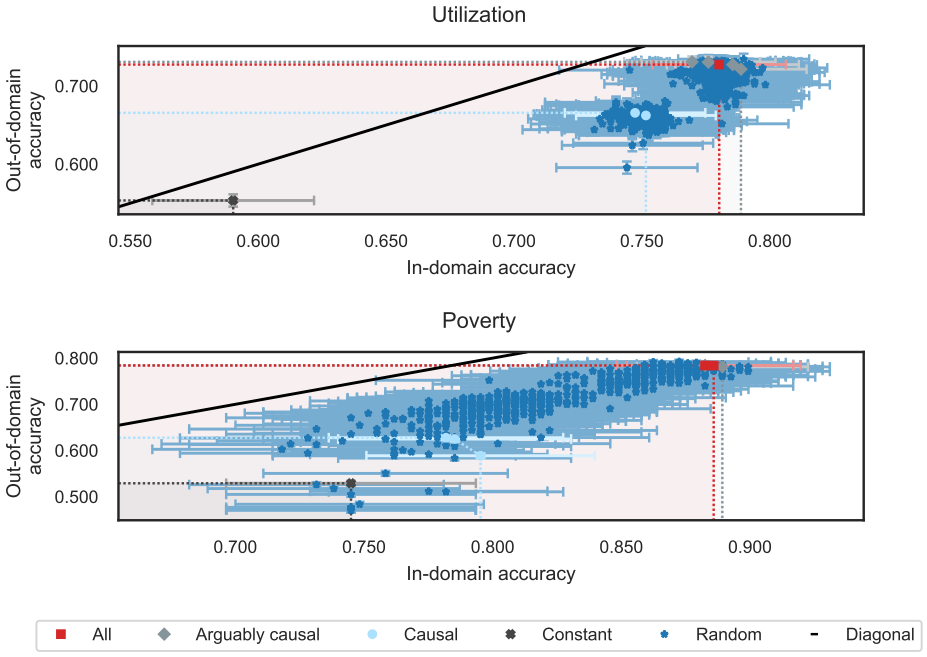


Figure 40: Performance of random subsets in comparison to causal feature selection and the whole feature set. (Continued)

C.6 Ablation of anti-causal and non-causal features

We conducted an ablation study and provided the results in Figure 41 to 45. We remove anti-causal and non-causal features one at a time and measure the corresponding out-of-domain accuracy. In the following, we discuss in detail the non-causal features whose removal significantly dropped the out-of-domain performance and try to give explanations. We split by task.

Food Stamps Target is food stamp recipiency in past year for households with child across geographic region.

- *Relationship to reference person:* There could be a stable and informative correlation within the survey of US Census between kind of household members (encoded in relationship to the reference person/head of the household, e.g., multiple generation household vs roommates) and food stamp recipiency. We didn't classify this variable as causal, as it is survey related.

Income Target is income level across geographic regions.

- *Relationship to reference person:* Same argument as in the task 'Food Stamps' applies.
- *Marital status:* Marital status and personal income are both intricately linked with socio-economic status, although we haven't found any research causally linking them together.
- *Insurance through a current or former employer or union / Medicare for people 65 or older, or people with certain disabilities:* These insurances are benefits not tied to income, but rather the person's employer or age and medical condition. They are however indicative of the economic and social environment of the individual, which is informative of the income level.
- *Year:* The year, e.g., 2018, encodes information about the economic status, which may be predictive across geographic regions.

Public Coverage Target is public coverage of non-Medicare eligible low-income individuals across disability status.

- *State / Year:* The current state of living and year encode information about the economic status.

Voting Target is whether an individual voted in the US presidential elections across geographic regions.

- *Party preference on specific topics, e.g. pollution / Opinion on party inclinations, e.g., which party favors stronger government / Opinion on sensitive topics, e.g., abortion, religion, gun control:* The opinions/preferences of an individual may sort them to specific sub-groups of the populations, wherein civil duty is or is not prominent. It is fathomable that similar sub-groups form across geographic regions.

Hypertension Target is high blood pressure across BMI categories.

- *State:* The current state of living encodes information about the socio-economic status, which research linked to hypertension in several studies [Leng et al., 2015].

Sepsis Target is sepsis across length of stay in ICU.

- *Hospital:* Hospitals serve different groups of the populations which differ in their risks of attaining sepsis.

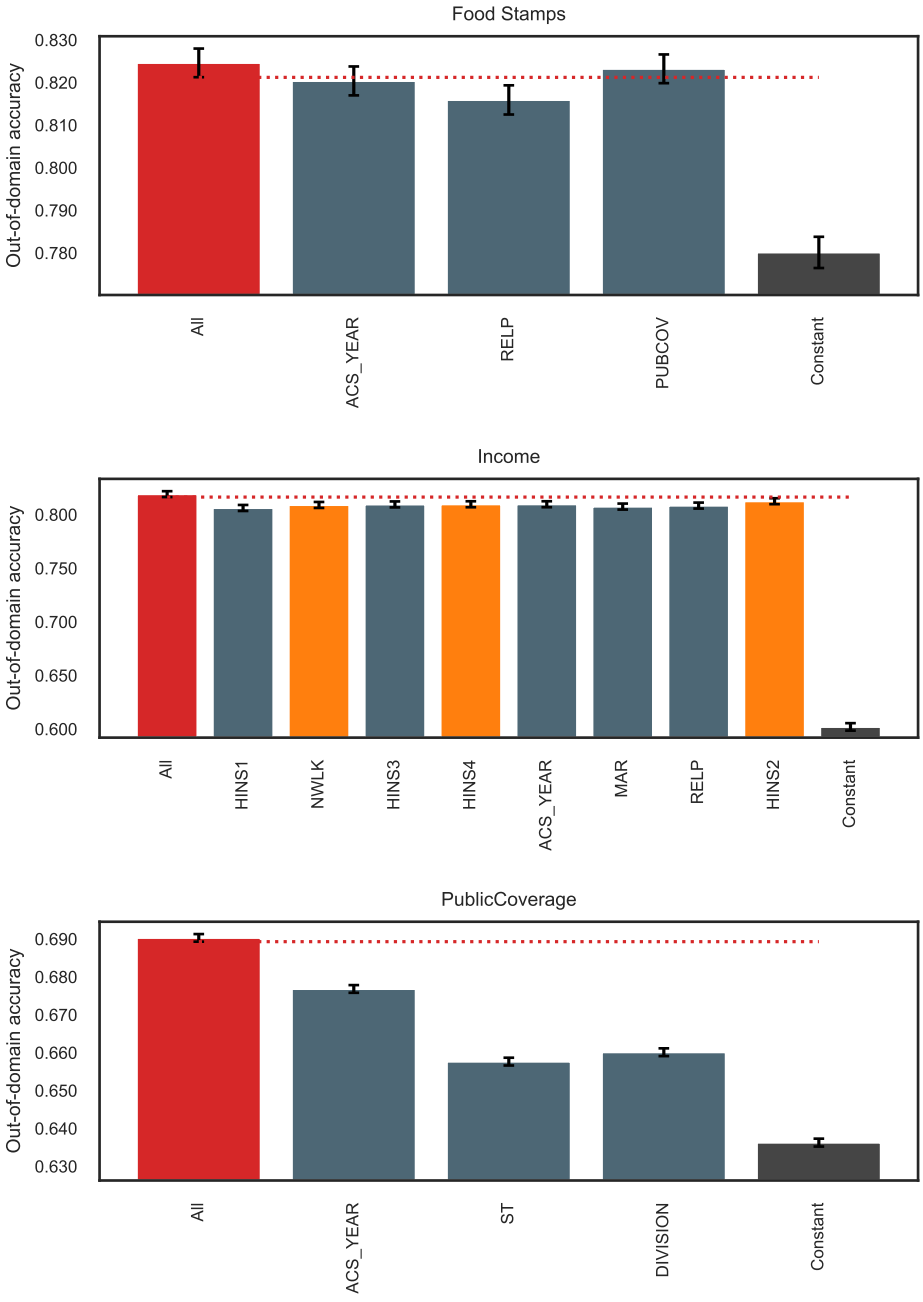


Figure 41: Removing one feature at a time. Anti-causal features are colored in orange, non-causal in grey.

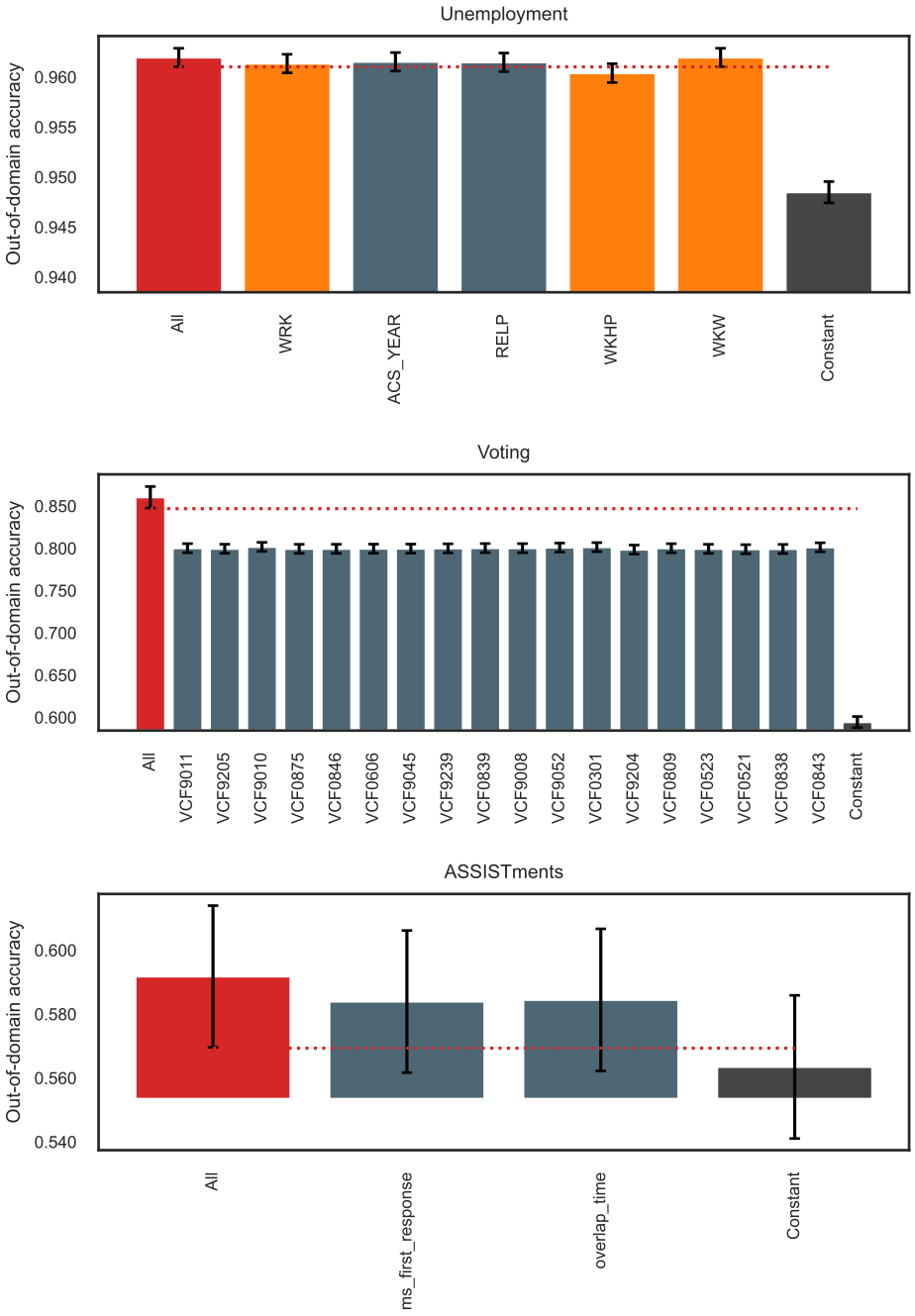


Figure 42: Removing one feature at a time. Anti-causal features are colored in orange, non-causal in grey. (Continued)

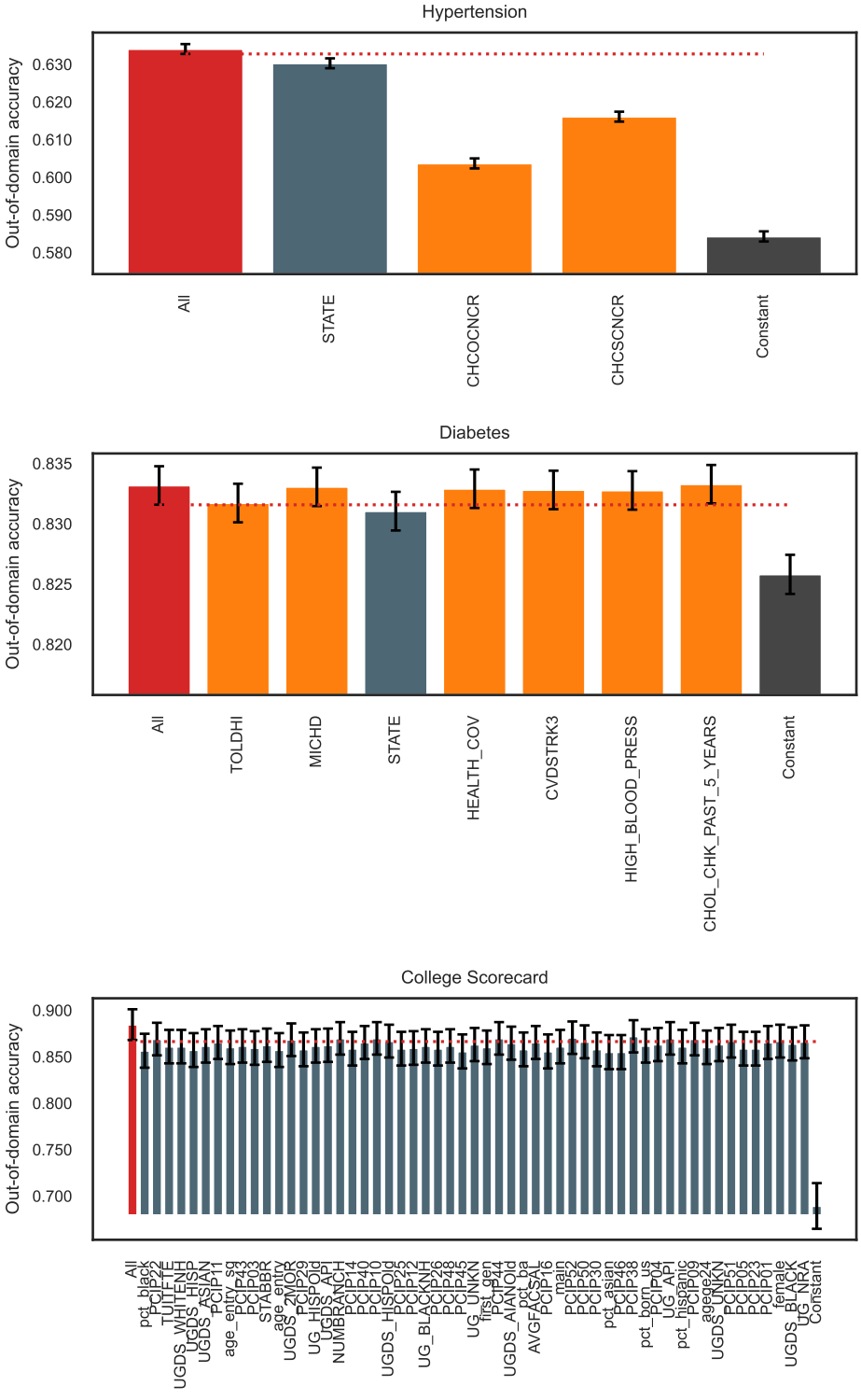


Figure 43: Removing one feature at a time. Anti-causal features are colored in orange, non-causal in grey. (Continued)

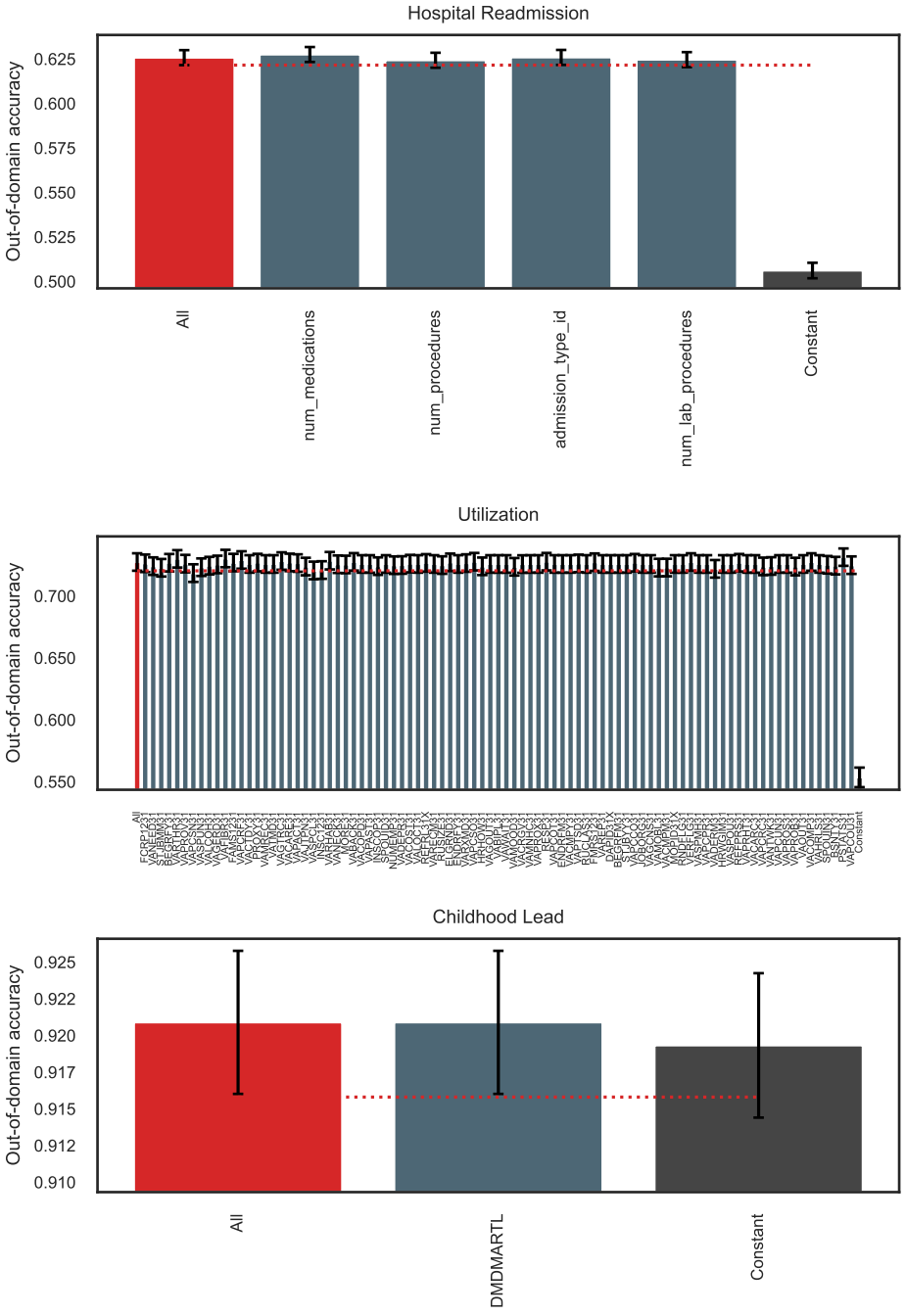


Figure 44: Removing one feature at a time. Anti-causal features are colored in orange, non-causal in grey. (Continued)

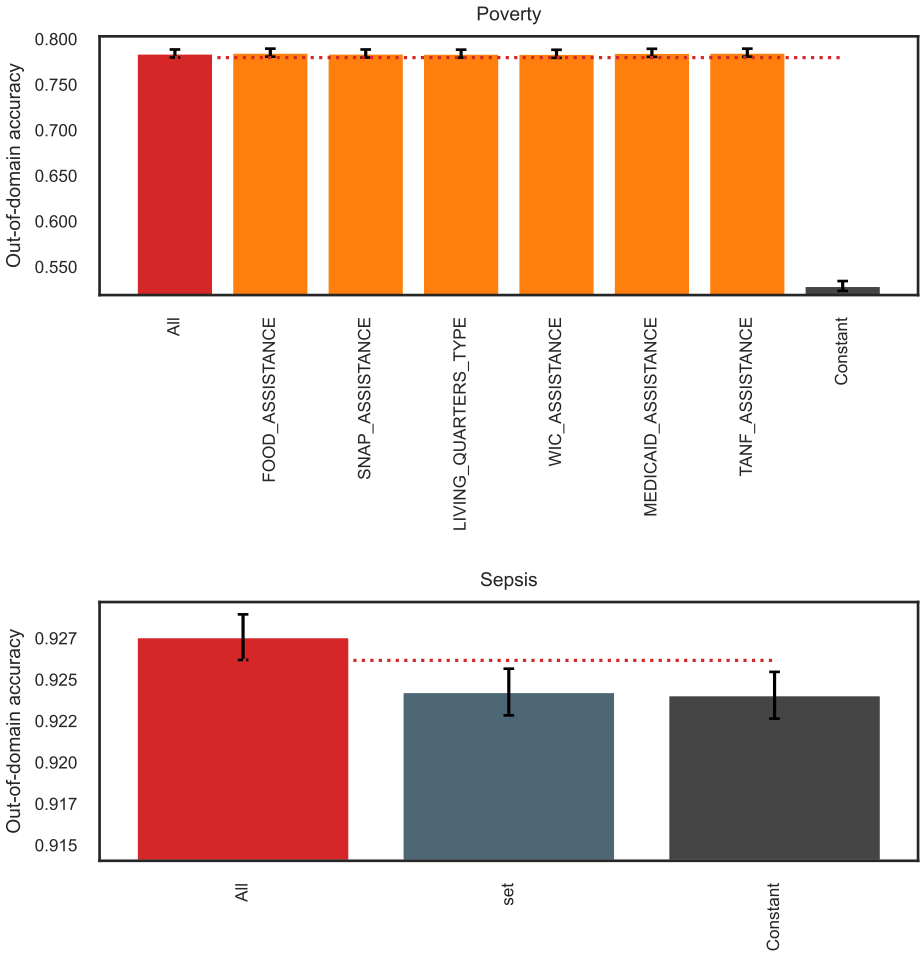


Figure 45: Removing one feature at a time. Anti-causal features are colored in orange, non-causal in grey. (Continued)

C.7 Empirical results across machine learning models

We show the Pareto-dominate performances for each machine learning model in Figure 46 - 49. The detailed results are provided at `https://github.com/socialfoundations/causal-features/tree/add-ons/experiments_causal/results/`. We have a summary table saved in a csv file for each task.

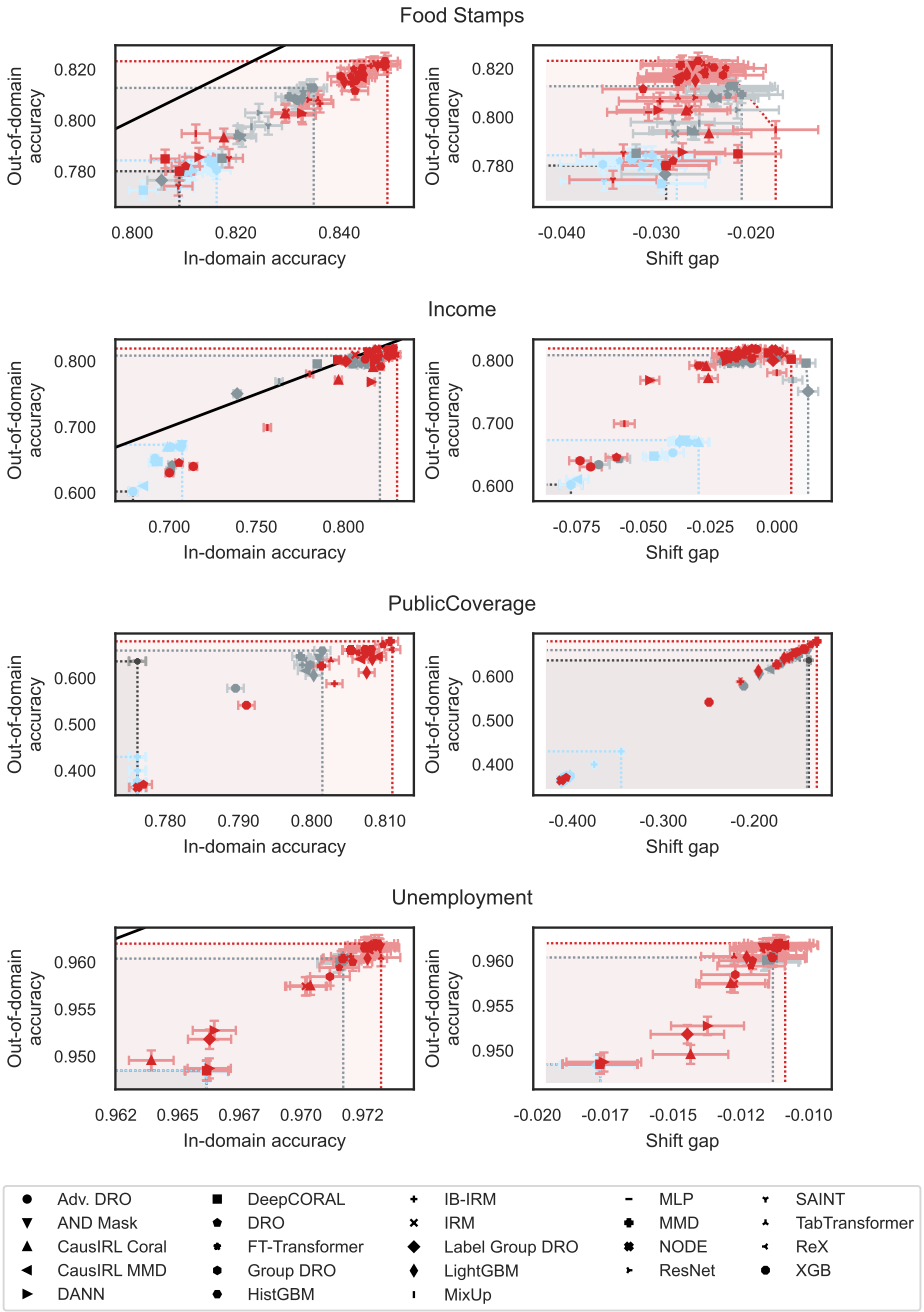


Figure 46: (Left) Pareto-dominate performance of in-domain and out-of-domain accuracy by feature selection and machine learning model. (Right) Pareto-dominate performance of shift gap and out-of-domain accuracy accomplished by feature selection and machine learning model. The feature sets are color-coded. Red indicates all features. The causal features are shown in blue, the arguably causal features in grey.

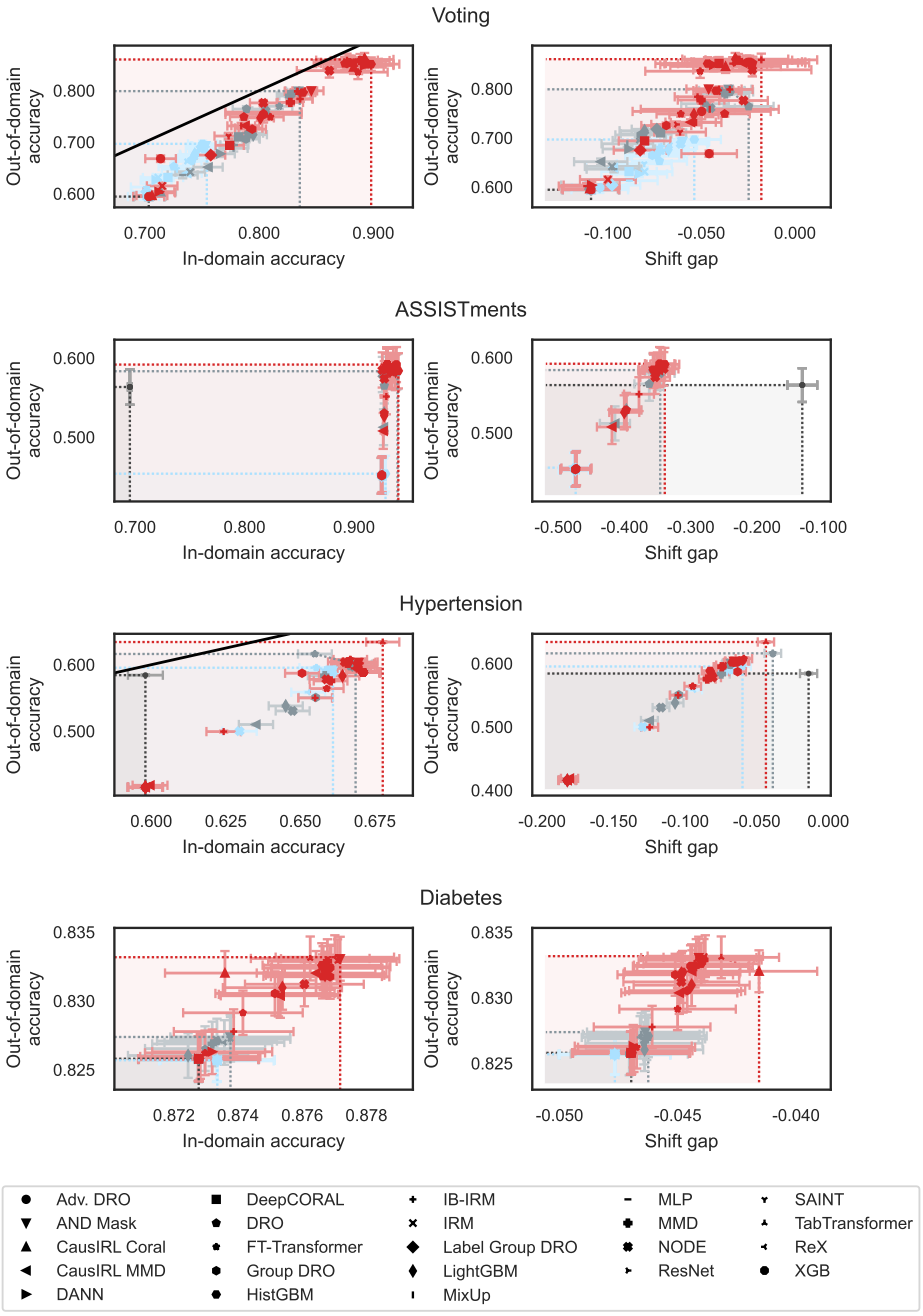


Figure 47: (Left) Pareto-dominate performance of in-domain and out-of-domain accuracy by feature selection and machine learning model. (Right) Pareto-dominate performance of shift gap and out-of-domain accuracy accomplished by feature selection and machine learning model. The feature sets are color-coded. Red indicates all features. The causal features are shown in blue, the arguably causal features in gray. (Continued)

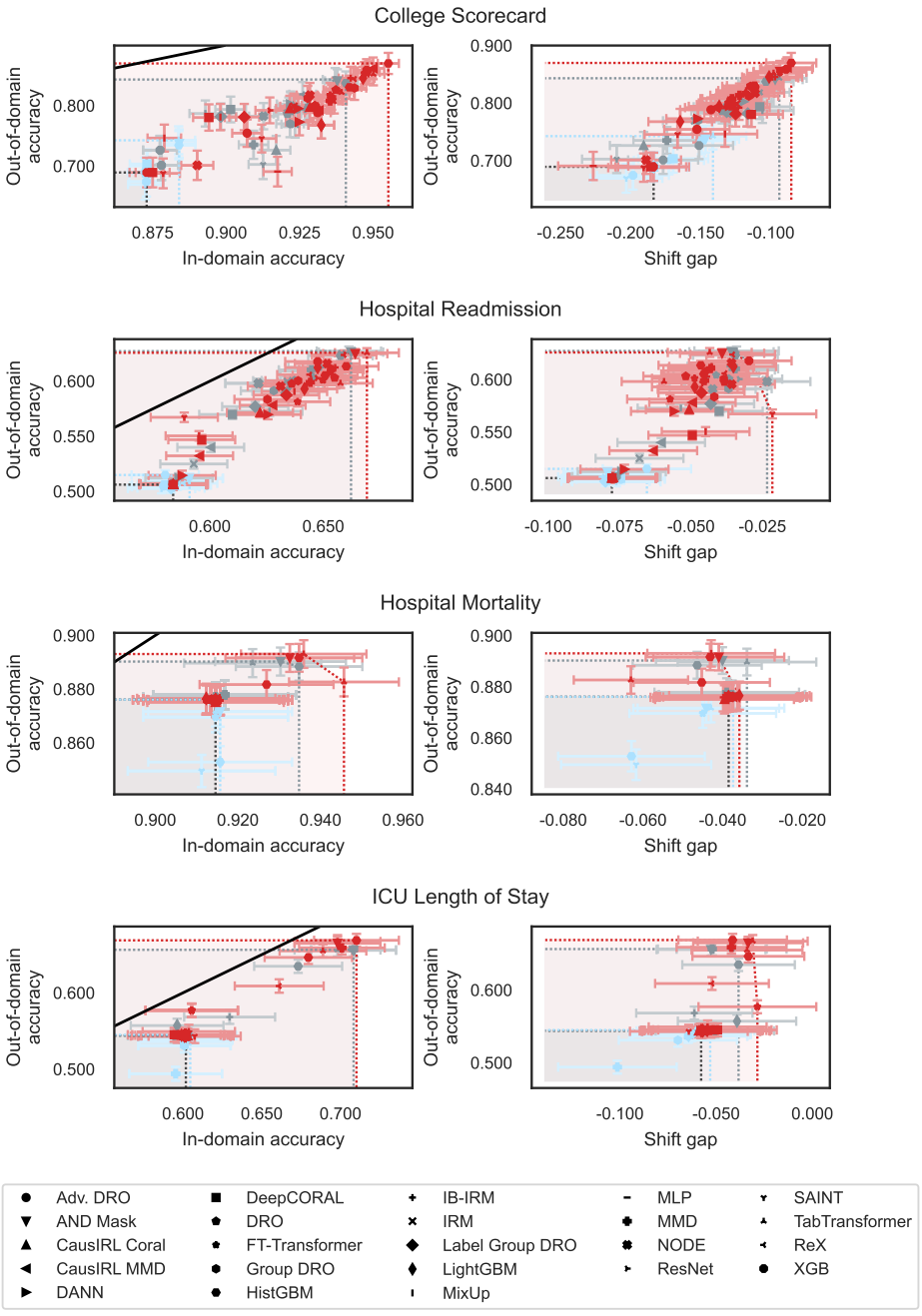


Figure 48: (Left) Pareto-dominate performance of in-domain and out-of-domain accuracy by feature selection and machine learning model. (Right) Pareto-dominate performance of shift gap and out-of-domain accuracy accomplished by feature selection and machine learning model. The feature sets are color-coded. Red indicates all features. The causal features are shown in blue, the arguably causal features in gray. (Continued)

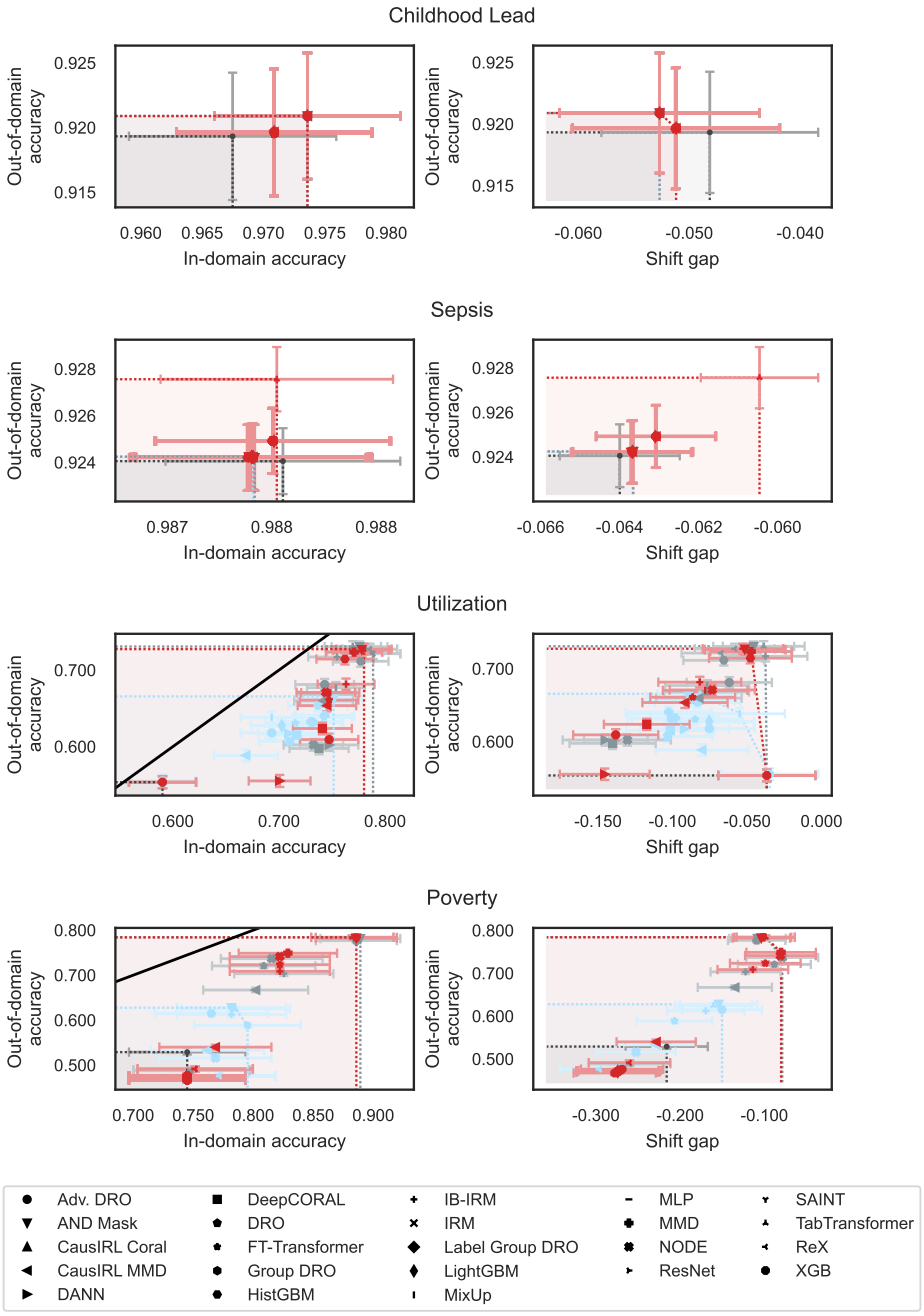


Figure 49: (Left) Pareto-dominate performance of in-domain and out-of-domain accuracy by feature selection and machine learning model. (Right) Pareto-dominate performance of shift gap and out-of-domain accuracy accomplished by feature selection and machine learning model. The feature sets are color-coded. Red indicates all features. The causal features are shown in blue, the arguably causal features in gray. (Continued)

D Synthetic experiments

We conducted synthetic experiments. The setup is depicted in Figure 50. The causal mechanisms are modeled as (i) linear with weights randomly drawn in (-1,1) and (ii) based on a neural network with random instantiation. The noise variables are drawn from a standard normal distribution. The task is to classify whether the target is larger than 0.

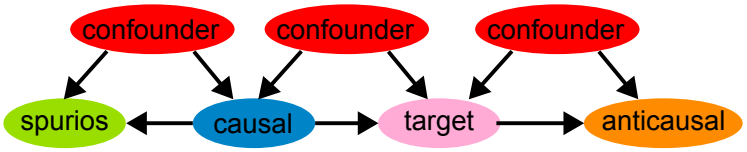
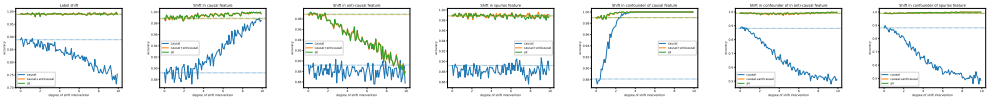
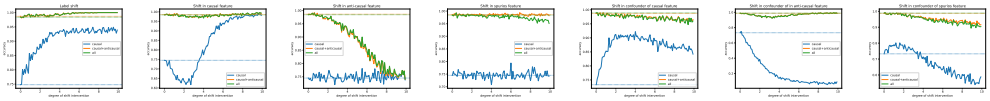


Figure 50: Causal graph to generate samples.

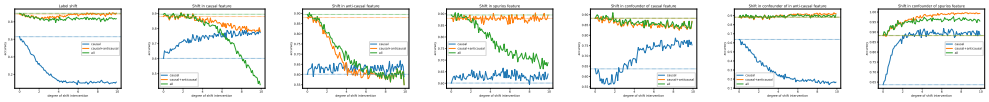
Similar to Rothenhäusler et al. [2020], we vary the degree of domain shift using shift intervention on target, features and confounders. We draw 1,000 training samples from the causal mechanism, and evaluate the performance on 1,000 testing samples from the intervented causal mechanism with shift interventions varying from 0 to 10; step size is 0.1. We provide example performances in Figure 51. Our code is based on the synthetic study conducted by Montagna et al. [2024].



(a) Linear mechanism and normal noise, estimated by logistic regression. Dotted lines indicate in-domain testing accuracy.



(b) Neural net mechanism and normal noise, estimated by logistic regression.



(c) Neural net mechanism and normal noise, estimated by neural net.

Figure 51: Synthetic experiments. Mechanisms are randomly instantiated. Task is to classify target > 0.

The synthetic experiments confirm our empirical findings. Using all features achieves best out-of-domain prediction accuracy. The one exception is if the distribution shift is exclusively on the anti-causal features and even in this case, a strong shift is needed before causal features achieve best out-of-domain accuracy.

E Tasks and data sources

In this section we give details on our tasks. We briefly describe the data source, target and distribution shift. We refer the reader to Gardner et al. [2023] for more details on the *Tableshift* tasks, their data sources and the reasoning behind the proposed domain shifts. We provide links to the datasets, their access and licenses in Table 7.

We list the features, and sort them into causal, arguably causal and anti-causal. In Section 2.1 and Appendix A, we justify the sorting for seven examples: 'Diabetes', 'Income', 'Unemployment', 'Hospital mortality', 'Hypertension', 'Voting' and 'ASSISTments'. While in good faith, we do the selection under epistemic uncertainty. Future research in health care and social science may rebut our sorting. Therefore, we advise caution when using our classification for follow-up research.

Table 8 provides an overview of all tasks, their training and testing domains, shift gap of the constant predictor and number of observation in the dataset. We provide additional insights into the distribution shift in Appendix E.12.

E.1 TableShift: ACS

We have multiple tasks based on American Community Survey (ACS) [U.S. Census Bureau, 2018], derived from *Folktables* [Ding et al., 2021]. The encoding is found in the ACS documentation.[9]

E.1.1 Foodstamps

Target: Food stamp recipiency in past year for households with child [FS]

Shift: Geographic region (U.S. divisions) [DIVISION]

List of causal features:
- Age in years [AGEP]
- Sex [SEX]
- Race [RAC1P]
- Place of birth [POBP]
- Disability [DIS]
- Hearing difficulty [DEAR]
- Vision difficulty [DEYE]
- Cognitive difficulty [DREM]
- Ancestry [ANC]
- Nativity [NATIVITY]
- Marital status [MAR]
- State [ST]

List of arguably causal features:
- Ability to speak English [ENG]
- Gave birth to child within the past 12 months [FER]
- Citizenship status [CIT]
- Educational attainment [SCHL]
- Households presence and age of children [HUPAC]
- Occupation [OCCP]
- Military service [MIL]
- Workers in family during the past 12 months [WIF]
- Usual hours worked per week past 12 months [WKHP]
- Weeks worked during past 12 months [WKW]
- Worked last week [WRK]
- On layoff from work [NWLA]
- Looking for work [NWLK]

List of other features:
- Year of survey [ACS_YEAR]
- Relationship to reference person [RELP]
- Public health coverage [PUBCOV]

E.1.2 Income

The selection procedure for the task 'Income' is discussed in detail in Appendix A.

Target: Total person's income $\geq$ 56k for employed adults [PINCP]

Shift: Geographic region (U.S. divisions) [DIVISION]

List of causal features:
- Age in years [AGEP]

[9] `https://www.census.gov/programs-surveys/acs/microdata/documentation.html`

- Sex [SEX]
- Race [RAC1P]
- Place of birth [POBP]

List of arguably causal features:
- State [ST]
- Ability to speak English [ENG]
- Gave birth to child within the past 12 months [FER]
- Citizenship status [CIT]
- Educational attainment [SCHL]
- Occupation [OCCP]
- Class of worker [COW]
- Usual hours worked per week past 12 months [WKHP]
- Weeks worked during past 12 months [WKW]
- Worked last week [WRK]
- On layoff from work [NWLA]

List of anti-causal features:
- Insurance purchased directly from an insurance company [HINS2]
- Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability [HINS4]
- Looking for work [NWLK]

List of other features:
- Year of survey [ACS_YEAR]
- Marital status [MAR]
- Insurance through a current or former employer or union [HINS1]
- Medicare, for people 65 and older, or people with certain disabilities [HINS3]
- Relationship to reference person [RELP]

E.1.3 Public Coverage

Target: Public health coverage [PUBCOV]

Shift: Disability status [DIS]

List of causal features:
- Age in years [AGEP]
- Sex [SEX]
- Race [RAC1P]
- Hearing difficulty [DEAR]
- Vision difficulty [DEYE]
- Cognitive difficulty [DREM]
- Ancestry [ANC]
- Nativity [NATIVITY]

List of arguably causal features:
- Employment status of parents [ESP]
- Total person's income in dollars [PINCP]
- Employment status [ESR]
- Gave birth to child within the past 12 months [FER]
- Marital status [MAR]
- Citizenship status [CIT]
- Educational attainment [SCHL]
- Mobility status [MIG]

List of other features:
- Year of survey [ACS_YEAR]
- State [ST]
- Geographic region [DIVISION]

E.1.4 Unemployment

The selection procedure for the task 'Unemployment' is discussed in detail in Appendix A.

Target: Employment status (is unemployed) [ESR]

Shift: Educational attainment [SCHL]

List of causal features:
- Age in years [AGEP]
- Sex [SEX]
- Race [RAC1P]
- Place of birth [POBP]
- Disability status [DIS]
- Ancestry [ANC]
- Nativity [NATIVITY]
- Hearing difficulty [DEAR]
- Vision difficulty [DEYE]
- Cognitive difficulty [DREM]
- Ambulatory difficulty [DPHY]

List of arguably causal features:
- Ability to speak English [ENG]
- Occupation [OCCP]
- Employment status of parents [ESP]
- Military service [MIL]
- Gave birth to child within the past 12 months [FER]
- Marital status [MAR]
- Citizenship status [CIT]
- Mobility status [MIG]
- State [ST]
- Geographic region [DIVISION]

List of anti-causal features:
- Usual hours worked per week past 12 months [WKHP]
- Weeks worked during past 12 months [WKW]
- Worked last week [WRK]

List of other features:
- Year of survey [ACS_YEAR]
- Relationship to reference person [RELP]

E.2 TableShift: ANES

We have one task based on American National Election Studies (ANES) [American National Election Studies, 2020].[10]

E.2.1 Voting

The selection procedure for the task 'Voting' is discussed in detail in Appendix A.

Target: Voted in national election [VCF0702]

Shift: Us census region [VCF0112]

List of causal features:
- Election year [VCF0004]
- State [VCF0901b]
- Registered to vote pre-election [VCF0701]
- Age [VCF0101]
- Gender [VCF0104]

[10] `https://electionstudies.org/`

- Race/ethnicity [VCF0105a]
- Occupation group [VCF0115]
- Education level [VCF0140a]

List of arguably causal features:
- Democratic party feeling thermometer [VCF0218]
- Republican party feeling thermometer [VCF0224]
- Party identification [VCF0302]
- Like-dislike scale placement for democratic party (0-10) [VCF9201]
- Like-dislike scale placement for republican party (0-10) [VCF9202]
- Do any of the parties in the U.S. represent views reasonably well [VCF9203]
- Better when one party controls both presidency and congress or when control is split [VCF9206]
- President thermometer [VCF0428]
- Vice-president thermometer [VCF0429]
- Rating of government economic policy [VCF0822]
- Better or worse economy in past year [VCF0870]
- Liberal-conservative scale [VCF0803]
- Approve participation in protests [VCF0601]
- Voting is the only way to have a say in government [VCF0612]
- It matters whether I vote [VCF0615]
- Those who don't care about election outcome should vote [VCF0616]
- Someone should vote if their party can't win [VCF0617]
- Interest in the elections [VCF0310]
- Belongs to political organization or club [VCF0743]
- Tried to influence others during campaign [VCF0717]
- Attended political meetings/rallies during campaign [VCF0718]
- Displayed candidate button/sticker during campaign [VCF0720]
- Donated money to party or candidate during campaign [VCF0721]
- How much of the time can you trust the media to report the news fairly [VCF0675]
- Watched tv programs about the election campaigns [VCF0724]
- Heard radio programs about the election campaigns [VCF0725]
- Read about the election campaigns in magazines [VCF0726]
- Saw election campaign information on the internet [VCF0745]

List of other features:
- Think of yourself as closer to the republican or democratic party [VCF0301]
- Party preference on pollution and environment [VCF9008]
- Party preference on inflation [VCF9010]
- Party preference on unemployment [VCF9011]
- Party in U.S. that represents views best [VCF9204]
- Which political party represents views best [VCF9205]
- Which party favors stronger government [VCF0521]
- Which party favors military spending cut [VCF0523]
- Most important national problem [VCF0875]
- Are things in U.S. going well or not [VCF9052]
- Guaranteed jobs and income scale (support/don't support) [VCF0809]
- Government services and spending scale (fewer/more services) [VCF0839]
- Defense spending scale (decrease/increase) [VCF0843]
- Position of the U.S. in past year [VCF9045]
- When should abortion be allowed by law [VCF0838]
- Importance of gun control [VCF9239]
- Importance of religion [VCF0846]
- How much does federal government waste tax money [VCF0606]

E.3 TableShift: BRFSS

We have two tasks based on Behavioral Risk Factor Surveillance System (BRFSS) [Centers for Disease Control and Prevention, 2021]. The encoding is found in BRFSS data dictionary.[11]

E.3.1 Diabetes

The selection procedure for the task 'Diabetes' is discussed in detail in Section 2.1.

Target: Diagnosed with diabetes [DIABETES]

Shift: Preferred race category [PRACE1]

List of causal features:
- Highest grade or year of school completed [EDUCA]
- Answer to the question 'Have you smoked at least 100 cigarettes in your entire life?' [SMOKE100]
- Sex of respondent [SEX]
- Marital status [MARITAL]

List of arguably causal features:
- Annual household income from all sources [INCOME]
- Number of days during the past 30 days where physical health was not good [PHYSHLTH]
- Body Mass Index (BMI) [BMI5]
- Body Mass Index (BMI) category [BMI5CAT]
- Answer to the question 'Do you now smoke cigarettes every day, some days, or not at all?' [SMOKDAY2]
- Consume Fruit 1 or more times per day [FRUIT_ONCE_PER_DAY]
- Consume vegetables 1 or more times per day [VEG_ONCE_PER_DAY]
- Total number of alcoholic beverages consumed per week [DRNK_PER_WEEK]
- Binge drinkers (males having five or more drinks on one occasion, females having four or more drinks on one occasion) [RFBING5]
- Physical activity or exercise during the past 30 days other than their regular job [TOTINDA]
- Time since last visit to the doctor for a checkup [CHECKUP1]
- Answer to the question 'Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?' [MEDCOST]
- Answer to the question 'for how many days during the past 30 days was your mental health not good?' [MENTHLTH]

List of anti-causal features:
- Diagnosed with high blood pressure [HIGH_BLOOD_PRESS]
- Time since last blooc cholesterol check [CHOL_CHK_PAST_5_YEARS]
- Diagnosed with high blood cholesterol [TOLDHI]
- Diagnosed past stroke [CVDSTRK3]
- Reports of coronary heart disease (CHD) or myocardial infarction (MI) [MICHD]
- Current health care coverage [HEALTH_COV]

List of other features:
- State [STATE]
- Year of BRFSS dataset [IYEAR]

E.3.2 Hypertension

The selection procedure for the task 'Hypertension' is discussed in detail in Appendix A.

Target: Diagnosed with high blood pressure [HIGH_BLOOD_PRESS]

Shift: Body Mass Index (BMI) category [BMI5CAT]

[11] `https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf`

List of causal features:
- Age group [AGEG5YR]
- Preferred race category [PRACE1]
- Sex of respondent [SEX]
- Answer to the question 'Have you smoked at least 100 cigarettes in your entire life?' [SMOKE100]
- Diagnosed with diabetes [DIABETES]

List of arguably causal features:
- Binary indicator for whether an individuals' income falls below the 2021 poverty guideline for family of four [POVERTY]
- Current employment status [EMPLOY1]
- Consume Fruit 1 or more times per day [FRUIT_ONCE_PER_DAY]
- Consume vegetables 1 or more times per day [VEG_ONCE_PER_DAY]
- Total number of alcoholic beverages consumed per week [DRNK_PER_WEEK]
- Binge drinkers (males having five or more drinks on one occasion, females having four or more drinks on one occasion) [RFBING5]
- Physical activity or exercise during the past 30 days other than their regular job [TOTINDA]
- Answer to the question 'Do you now smoke cigarettes every day, some days, or not at all?' [SMOKDAY2]
- Answer to the question 'Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?' [MEDCOST]

List of anti-causal features:
- Diagnosed with skin cancer [CHCSCNCR]
- Diagnosed with any other types of cancer [CHCOCNCR]

List of other features:
- State [STATE]
- Year of BRFSS dataset [IYEAR]

E.4 TableShift: ED

We have one task based on the data that appear on the college scorecard by the U.S. Department of Education (ED) [U.S. Department of Education, 2023].[12]

E.4.1 College Scorecard

Target: Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion/6 years) [C150_4]

Shift: Carnegie Classification - basic [CCBASIC]

List of causal features:
- Accreditor for institution [AccredAgency]
- Highest degree awarded [HIGHDEG]
- Control of institution [CONTROL]
- Region (IPEDS) [region]
- Locale of institution [LOCALE]
- Degree of urbanization of institution [locale2]
- Flag for Historically Black College and University [HBCU]
- Flag for distance-education-only education [DISTANCEONLY]
- Poverty rate, via Census data [poverty_rate]
- Unemployment rate, via Census data [unemp_rate]
- Carnegie Classification - size and setting [CCSIZSET]

List of arguably causal features:
- In-state tuition and fees [TUITIONFEE_IN]
- Out-of-state tuition and fees [TUITIONFEE_OUT]
- Tuition and fees for program-year institutions [TUITIONFEE_PROG]

[12] `https://collegescorecard.ed.gov/`

- Admission rate [ADM_RATE]
- Admission rate for all campuses rolled up to the 6-digit OPE ID [ADM_RATE_ALL]
- Midpoint of SAT scores at the institution (critical reading) [SATVRMID]
- Midpoint of SAT scores at the institution (math) [SATMTMID]
- Midpoint of SAT scores at the institution (writing) [SATWRMID]
- Midpoint of the ACT cumulative score [ACTCMMID]
- Midpoint of the ACT English score [ACTENMID]
- Midpoint of the ACT math score [ACTMTMID]
- Midpoint of the ACT writing score [ACTWRMID]
- Average net price for the largest program at the institution for program-year institutions [NPT4_PROG]
- Average cost of attendance (academic year institutions) [COSTT4_A]
- Average cost of attendance (program-year institutions) [COSTT4_P]
- Share of students who received a federal loan while in school [loan_ever]
- Share of students who received a Pell Grant while in school [pell_ever]
- Percentage of undergraduates who receive a Pell Grant [PCTPELL]
- Median household income [median_hh_inc]
- Average family income [faminc]
- Median family income [md_faminc]
- Enrollment of undergraduate degree-seeking students [UGDS]
- Enrollment of all undergraduate students [UG]

List of other features:

- State postcode [STABBR]
- Predominant degree awarded (recoded 0s and 4s) [sch_deg]
- Flag for main campus [main]
- Number of branch campuses [NUMBRANCH]
- Percentage of degrees awarded in Agriculture, Agriculture Operations, And Related Sciences [PCIP01]
- Percentage of degrees awarded in Natural Resources And Conservation [PCIP03]
- Percentage of degrees awarded in Architecture And Related Services [PCIP04]
- Percentage of degrees awarded in Area, Ethnic, Cultural, Gender, And Group Studies [PCIP05]
- Percentage of degrees awarded in Communication, Journalism, And Related Programs [PCIP09]
- Percentage of degrees awarded in Communications Technologies/Technicians And Support Services [PCIP10]
- Percentage of degrees awarded in Computer And Information Sciences And Support Services [PCIP11]
- Percentage of degrees awarded in Personal And Culinary Services [PCIP12]
- Percentage of degrees awarded in Education [PCIP13]
- Percentage of degrees awarded in Engineering [PCIP14]
- Percentage of degrees awarded in Engineering Technologies And Engineering-Related Fields [PCIP15]
- Percentage of degrees awarded in Foreign Languages, Literatures, And Linguistics [PCIP16]
- Percentage of degrees awarded in Family And Consumer Sciences/Human Sciences [PCIP19]
- Percentage of degrees awarded in Legal Professions And Studies [PCIP22]
- Percentage of degrees awarded in English Language And Literature/Letters [PCIP23]
- Percentage of degrees awarded in Liberal Arts And Sciences, General Studies And Humanities [PCIP24]
- Percentage of degrees awarded in Library Science [PCIP25]
- Percentage of degrees awarded in Biological And Biomedical Sciences [PCIP26]

- Percentage of degrees awarded in Mathematics And Statistics [PCIP27]
- Percentage of degrees awarded in Military Technologies And Applied Sciences [PCIP29]
- Percentage of degrees awarded in Multi/Interdisciplinary Studies [PCIP30]
- Percentage of degrees awarded in Parks, Recreation, Leisure, And Fitness Studies [PCIP31]
- Percentage of degrees awarded in Philosophy And Religious Studies [PCIP38]
- Percentage of degrees awarded in Theology And Religious Vocations [PCIP39]
- Percentage of degrees awarded in Physical Sciences [PCIP40]
- Percentage of degrees awarded in Science Technologies/Technicians [PCIP41]
- Percentage of degrees awarded in Psychology [PCIP42]
- Percentage of degrees awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services [PCIP43]
- Percentage of degrees awarded in Public Administration And Social Service Professions [PCIP44]
- Percentage of degrees awarded in Social Sciences [PCIP45]
- Percentage of degrees awarded in Construction Trades [PCIP46]
- Percentage of degrees awarded in Mechanic And Repair Technologies/Technicians [PCIP47]
- Percentage of degrees awarded in Precision Production [PCIP48]
- Percentage of degrees awarded in Transportation And Materials Moving [PCIP49]
- Percentage of degrees awarded in Visual And Performing Arts [PCIP50]
- Percentage of degrees awarded in Health Professions And Related Programs [PCIP51]
- Percentage of degrees awarded in Business, Management, Marketing, And Related Support Services [PCIP52]
- Percentage of degrees awarded in History [PCIP54]
- Total share of enrollment of undergraduate degree-seeking students who are white [UGDS_WHITE]
- Total share of enrollment of undergraduate degree-seeking students who are black [UGDS_BLACK]
- Total share of enrollment of undergraduate degree-seeking students who are Hispanic [UGDS_HISP]
- Total share of enrollment of undergraduate degree-seeking students who are Asian [UGDS_ASIAN]
- Total share of enrollment of undergraduate degree-seeking students who are American Indian/Alaska Native [UGDS_AIAN]
- Total share of enrollment of undergraduate degree-seeking students who are Native Hawaiian/Pacific Islander [UGDS_NHPI]
- Total share of enrollment of undergraduate degree-seeking students who are two or more races [UGDS_2MOR]
- Total share of enrollment of undergraduate degree-seeking students who are non-resident aliens [UGDS_NRA]
- Total share of enrollment of undergraduate degree-seeking students whose race is unknown [UGDS_UNKN]
- Total share of enrollment of undergraduate degree-seeking students who are white non-Hispanic [UGDS_WHITENH]
- Total share of enrollment of undergraduate degree-seeking students who are black non-Hispanic [UGDS_BLACKNH]
- Total share of enrollment of undergraduate degree-seeking students who are Asian/Pacific Islander [UGDS_API]
- Total share of enrollment of undergraduate degree-seeking students who are American Indian/Alaska Native [UGDS_AIANOld]
- Total share of enrollment of undergraduate degree-seeking students who are Hispanic [UGDS_HISPOld]

- Total share of enrollment of undergraduate students who are non-resident aliens [UG_NRA]
- Total share of enrollment of undergraduate students whose race is unknown [UG_UNKN]
- Total share of enrollment of undergraduate students who are white non-Hispanic [UG_WHITENH]
- Total share of enrollment of undergraduate students who are black non-Hispanic [UG_BLACKNH]
- Total share of enrollment of undergraduate students who are Asian/Pacific Islander [UG_API]
- Total share of enrollment of undergraduate students who are American Indian/Alaska Native [UG_AIANOld]
- Total share of enrollment of undergraduate students who are Hispanic [UG_HISPOld]
- Share of undergraduate, degree-/certificate-seeking students who are part-time [PPTUG_EF]
- Share of undergraduate, degree-/certificate-seeking students who are part-time [PPTUG_EF2]
- Net tuition revenue per full-time equivalent student [TUITFTE]
- Instructional expenditures per full-time equivalent student [INEXPFTE]
- Average faculty salary [AVGFACSAL]
- Proportion of faculty that is full-time [PFTFAC]
- Average age of entry, via SSA data [age_entry]
- Average of the age of entry squared [age_entry_sq]
- Percent of students over 23 at entry [agege24]
- Share of female students, via SSA data [female]
- Share of married students [married]
- Share of dependent students [dependent]
- Share of veteran students [veteran]
- Share of first-generation students [first_gen]
- Percent of the population from students' zip codes that is White, via Census data [pct_white]
- Percent of the population from students' zip codes that is Black, via Census data [pct_black]
- Percent of the population from students' zip codes that is Asian, via Census data [pct_asian]
- Percent of the population from students' zip codes that is Hispanic, via Census data [pct_hispanic]
- Percent of the population from students' zip codes with a bachelor's degree over the age 25, via Census data [pct_ba]
- Percent of the population from students' zip codes over 25 with a professional degree, via Census data [pct_grad_prof]
- Percent of the population from students' zip codes that was born in the US, via Census data [pct_born_us]

E.5 TableShift: Kaggle

We have one task based the data collected from an online learning tool[13] and released on Kaggle [Feng et al., 2009].

E.5.1 ASSISTments

The selection procedure for the task 'ASSISTments' is discussed in detail in Appendix A.

[13] `https://new.assistments.org/`

Target: Correct on first attempt [correct]

Shift: School [school_id]

List of causal features:
- Number of hints on this problem. [hint_count]
- Number of student attempts on this problem. [attempt_count]
- ID of the skill associated with the problem [skill_id]
- Problem type [problem_type]
- Whether or not the student asks for all hints [bottom_hint]
- Tutor/Test mode [tutor_mode]
- Assignment position on the class assignments page [position]
- Type of the head section of the problem set[type]
- Type of first action: attempt or ask for a hint [first_action]

List of arguably causal features:
- Predicted Boredom of student for the problem [Average_confidence(BORED)]
- Predicted Engaged Concentration of student for the problem [Average_confidence(CONCENTRATING)]
- Predicted Confusion of student for the problem [Average_confidence(CONFUSED)]
- Predicted Frustration of student for the problem [Average_confidence(FRUSTRATED)]

List of other features:
- Time in milliseconds for the student's first response [ms_first_response]
- Time in milliseconds for the student's overlap time [overlap_time]

E.6 TableShift: MIMIC

We have two tasks based on Medical Information Mart for Intensive Care (MIMIC-III), derived from MIMIC-Extract [Johnson et al., 2016, Wang et al., 2020a,b].

E.6.1 Stay in ICU

Target: Stay in ICU for longer than 3 days [los_3]

Shift: Insurance type (Medicare, Private, Medicaid, Government, Self Pay) [insurance]

List of causal features:
- Age in years [age]
- Gender [gender]
- Ethnicity [ethnicity]
- Height [height_mean_0]
- Weight [weight_mean_0]

List of arguably causal features:
- Bicarbonate [bicarbonate_mask_0, . . . , bicarbonate_mask_23, bicarbonate_mean_0, . . . , bicarbonate_mean_23, bicarbonate_time_since_measured_0, . . . , bicarbonate_time_since_measured_23]
- Co2 [co2_mask_0, . . . , co2_mask_23, co2_mean_0, . . . , co2_mean_23, co2_time_since_measured_0, . . . , co2_time_since_measured_23]
- Partial pressure of carbon dioxide (pCO2) and end_tidal CO2 (ETCO2) [co2_(etco2_pco2_etc)_mask_0, . . . , co2_(etco2_pco2_etc)_mask_23, co2_(etco2_pco2_etc)_mean_0, . . . , co2_(etco2_pco2_etc)_mean_23, co2_(etco2_pco2_etc)_time_since_measured_0, . . . , co2_(etco2_pco2_etc)_time_since_measured_23]
- Partial pressure of oxygen [partial_pressure_of_oxygen_mask_0, . . . , partial_pressure_of_oxygen_mask_23, partial_pressure_of_oxygen_mean_0, . . . , partial_pressure_of_oxygen_mean_23, partial_pressure_of_oxygen_time_since_measured_0, . . . , partial_pressure_of_oxygen_time_since_measured_23]

- Fraction inspired oxygen [fraction_inspired_oxygen_mask_0, . . . , fraction_inspired_oxygen_mask_23, fraction_inspired_oxygen_mean_0, . . . , fraction_inspired_oxygen_mean_23, fraction_inspired_oxygen_time_since_measured_0, . . . , fraction_inspired_oxygen_time_since_measured_23, fraction_inspired_oxygen_set_mask_0, . . . , fraction_inspired_oxygen_set_mask_23, fraction_inspired_oxygen_set_mean_0, . . . , fraction_inspired_oxygen_set_mean_23, fraction_inspired_oxygen_set_time_since_measured_0, . . . , fraction_inspired_oxygen_set_time_since_measured_23]
- Glascow coma score [glascow_coma_scale_total_mask_0, . . . , glascow_coma_scale_total_mask_23, glascow_coma_scale_total_mean_0, . . . , glascow_coma_scale_total_mean_23, glascow_coma_scale_total_time_since_measured_0, . . . , glascow_coma_scale_total_time_since_measured_23]
- Lactate [lactate_mask_0, . . . , lactate_mask_23, lactate_mean_0, . . . , lactate_mean_23, lactate_time_since_measured_0, . . . , lactate_time_since_measured_23]
- Lactic acid [lactic_acid_mask_0, . . . , lactic_acid_mask_23, lactic_acid_mean_0, . . . , lactic_acid_mean_23, lactic_acid_time_since_measured_0, . . . , lactic_acid_time_since_measured_23]
- Sodium [sodium_mask_0, . . . , sodium_mask_23, sodium_mean_0, . . . , sodium_mean_23, sodium_time_since_measured_0, . . . , sodium_time_since_measured_23]
- Hemoglobin [hemoglobin_mask_0, . . . , hemoglobin_mask_23, hemoglobin_mean_0, . . . , hemoglobin_mean_23, hemoglobin_time_since_measured_0, . . . , hemoglobin_time_since_measured_23]
- Mean blood pressure [mean_blood_pressure_mask_0, . . . , mean_blood_pressure_mask_23, mean_blood_pressure_mean_0, . . . , mean_blood_pressure_mean_23, mean_blood_pressure_time_since_measured_0, . . . , mean_blood_pressure_time_since_measured_23]
- Oxygen saturation [oxygen_saturation_mask_0, . . . , oxygen_saturation_mask_23, oxygen_saturation_mean_0, . . . , oxygen_saturation_mean_23, oxygen_saturation_time_since_measured_0, . . . , oxygen_saturation_time_since_measured_23]
- Ph [ph_mask_0, . . . , ph_mask_23, ph_mean_0, . . . , ph_mean_23, ph_time_since_measured_0, . . . , ph_time_since_measured_23]
- Respiratory rate [respiratory_rate_mask_0, . . . , respiratory_rate_mask_23, respiratory_rate_mean_0, . . . , respiratory_rate_mean_23, respiratory_rate_time_since_measured_0, . . . , respiratory_rate_time_since_measured_23, respiratory_rate_set_mask_0, . . . , respiratory_rate_set_mask_23, respiratory_rate_set_mean_0, . . . , respiratory_rate_set_mean_23, respiratory_rate_set_time_since_measured_0, . . . , respiratory_rate_set_time_since_measured_23]
- Systolic blood pressure [systolic_blood_pressure_mask_0, . . . , systolic_blood_pressure_mask_23, systolic_blood_pressure_mean_0, . . . , systolic_blood_pressure_mean_23, systolic_blood_pressure_time_since_measured_0, . . . , systolic_blood_pressure_time_since_measured_23]
- Heart rate [heart_rate_mask_0, . . . , heart_rate_mask_23, heart_rate_mean_0, . . . , heart_rate_mean_23, heart_rate_time_since_measured_0, . . . , heart_rate_time_since_measured_23]
- Temperature [temperature_mask_0, . . . , temperature_mask_23, temperature_mean_0, . . . , temperature_mean_23, temperature_time_since_measured_0, . . . , temperature_time_since_measured_23]
- White blood cell count[white_blood_cell_count_mask_0, . . . , white_blood_cell_count_mask_23, white_blood_cell_count_mean_0, . . . , white_blood_cell_count_mean_23, white_blood_cell_count_time_since_measured_0, . . . , white_blood_cell_count_time_since_measured_23]

List of other features:

- Height [height_mask_0, . . . , height_mask_23, height_mean_1, . . . , height_mean_23, height_time_since_measured_0, . . . , height_time_since_measured_23]
- Weight [weight_mask_0, . . . , weight_mask_23, weight_mean_1, . . . , weight_mean_23, weight_time_since_measured_0, . . . , weight_time_since_measured_23]
- Alanine aminotransferase [alanine_aminotransferase_mask_0, . . . , alanine_aminotransferase_mask_23, alanine_aminotransferase_mean_0, . . . , alanine_aminotransferase_mean_23, alanine_aminotransferase_time_since_measured_0, . . . , alanine_aminotransferase_time_since_measured_23]
- Albumin [albumin_mask_0, . . . , albumin_mask_23, albumin_mean_0, . . . , albumin_mean_23, albumin_time_since_measured_0, . . . , albumin_time_since_measured_23]
- Alanine aminotransferase [albumin_ascites_mask_0, . . . , albumin_ascites_mask_23, albumin_ascites_mean_0, . . . , albumin_ascites_mean_23, albumin_ascites_time_since_measured_0, . . . , albumin_ascites_time_since_measured_23]
- Albumin pleural [albumin_pleural_mask_0, . . . , albumin_pleural_mask_23, albumin_pleural_mean_0, . . . , albumin_pleural_mean_23, albumin_pleural_time_since_measured_0, . . . , albumin_pleural_time_since_measured_23]
- Albumin in urine [albumin_urine_mask_0, . . . , albumin_urine_mask_23, albumin_urine_mean_0, . . . , albumin_urine_mean_23, albumin_urine_time_since_measured_0, . . . , albumin_urine_time_since_measured_23]
- Alkaline phosphate [alkaline_phosphate_mask_0, . . . , alkaline_phosphate_mask_23, alkaline_phosphate_mean_0, . . . , alkaline_phosphate_mean_23, alkaline_phosphate_time_since_measured_0, . . . , alkaline_phosphate_time_since_measured_23]
- Anion gap [anion_gap_mask_0, . . . , anion_gap_mask_23, anion_gap_mean_0, . . . , anion_gap_mean_23, anion_gap_time_since_measured_0, . . . , anion_gap_time_since_measured_23]
- Asparate aminotransferase [asparate_aminotransferase_mask_0, . . . , asparate_aminotransferase_mask_23, asparate_aminotransferase_mean_0, . . . , asparate_aminotransferase_mean_23, asparate_aminotransferase_time_since_measured_0, . . . , asparate_aminotransferase_time_since_measured_23]
- Basophils [basophils_mask_0, . . . , basophils_mask_23, basophils_mean_0, . . . , basophils_mean_23, basophils_time_since_measured_0, . . . , basophils_time_since_measured_23]
- Bilirubin [bilirubin_mask_0, . . . , bilirubin_mask_23, bilirubin_mean_0, . . . , bilirubin_mean_23, bilirubin_time_since_measured_0, . . . , bilirubin_time_since_measured_23]
- Blood urea nitrogen [blood_urea_nitrogen_mask_0, . . . , blood_urea_nitrogen_mask_23, blood_urea_nitrogen_mean_0, . . . , blood_urea_nitrogen_mean_23, blood_urea_nitrogen_time_since_measured_0, . . . , blood_urea_nitrogen_time_since_measured_23]
- Calcium [calcium_mask_0, . . . , calcium_mask_23, calcium_mean_0, . . . , calcium_mean_23, calcium_time_since_measured_0, . . . , calcium_time_since_measured_23]
- Calcium ionized [calcium_ionized_mask_0, . . . , calcium_ionized_mask_23, calcium_ionized_mean_0, . . . , calcium_ionized_mean_23, calcium_ionized_time_since_measured_0, . . . , calcium_ionized_time_since_measured_23]
- Calcium in urine [calcium_urine_mask_0, . . . , calcium_urine_mask_23, calcium_urine_mean_0, . . . , calcium_urine_mean_23, calcium_urine_time_since_measured_0, . . . , calcium_urine_time_since_measured_23]

- Cardiac index [cardiac_index_mask_0, . . . , cardiac_index_mask_23, cardiac_index_mean_0, . . . , cardiac_index_mean_23, cardiac_index_time_since_measured_0, . . . , cardiac_index_time_since_measured_23]
- Cardiac output by Fick principle [cardiac_output_fick_mask_0, . . . , cardiac_output_fick_mask_23, cardiac_output_fick_mean_0, . . . , cardiac_output_fick_mean_23, cardiac_output_fick_time_since_measured_0, . . . , cardiac_output_fick_time_since_measured_23]
- Cardiac output by thermodilution [cardiac_output_thermodilution_mask_0, . . . , cardiac_output_thermodilution_mask_23, cardiac_output_thermodilution_mean_0, . . . , cardiac_output_thermodilution_mean_23, cardiac_output_thermodilution_time_since_measured_0, . . . , cardiac_output_thermodilution_time_since_measured_23]
- Central venous pressure [central_venous_pressure_mask_0, . . . , central_venous_pressure_mask_23, central_venous_pressure_mean_0, . . . , central_venous_pressure_mean_23, central_venous_pressure_time_since_measured_0, . . . , central_venous_pressure_time_since_measured_23]
- Chloride [chloride_mask_0, . . . , chloride_mask_23, chloride_mean_0, . . . , chloride_mean_23, chloride_time_since_measured_0, . . . , chloride_time_since_measured_23]
- Chloride in urine [chloride_urine_mask_0, . . . , chloride_urine_mask_23, chloride_urine_mean_0, . . . , chloride_urine_mean_23, chloride_urine_time_since_measured_0, . . . , chloride_urine_time_since_measured_23]
- Cholesterol [cholesterol_mask_0, . . . , cholesterol_mask_23, cholesterol_mean_0, . . . , cholesterol_mean_23, cholesterol_time_since_measured_0, . . . , cholesterol_time_since_measured_23]
- HDL cholesterol [cholesterol_hdl_mask_0, . . . , cholesterol_hdl_mask_23, cholesterol_hdl_mean_0, . . . , cholesterol_hdl_mean_23, cholesterol_hdl_time_since_measured_0, . . . , cholesterol_hdl_time_since_measured_23]
- LDL cholesterol [cholesterol_ldl_mask_0, . . . , cholesterol_ldl_mask_23, cholesterol_ldl_mean_0, . . . , cholesterol_ldl_mean_23, cholesterol_ldl_time_since_measured_0, . . . , cholesterol_ldl_time_since_measured_23]
- Creatinine [creatinine_mask_0, . . . , creatinine_mask_23, creatinine_mean_0, . . . , creatinine_mean_23, creatinine_time_since_measured_0, . . . , creatinine_time_since_measured_23]
- Creatinine ascites [creatinine_ascites_mask_0, . . . , creatinine_ascites_mask_23, creatinine_ascites_mean_0, . . . , creatinine_ascites_mean_23, creatinine_ascites_time_since_measured_0, . . . , creatinine_ascites_time_since_measured_23]
- Creatinine body fluid [creatinine_body_fluid_mask_0, . . . , creatinine_body_fluid_mask_23, creatinine_body_fluid_mean_0, . . . , creatinine_body_fluid_mean_23, creatinine_body_fluid_time_since_measured_0, . . . , creatinine_body_fluid_time_since_measured_23]
- Creatinine pleural [creatinine_pleural_mask_0, . . . , creatinine_pleural_mask_23, creatinine_pleural_mean_0, . . . , creatinine_pleural_mean_23, creatinine_pleural_time_since_measured_0, . . . , creatinine_pleural_time_since_measured_23]
- Creatinine in urine [creatinine_urine_mask_0, . . . , creatinine_urine_mask_23, creatinine_urine_mean_0, . . . , creatinine_urine_mean_23, creatinine_urine_time_since_measured_0, . . . , creatinine_urine_time_since_measured_23]
- Diastolic blood pressure [diastolic_blood_pressure_mask_0, . . . , diastolic_blood_pressure_mask_23, diastolic_blood_pressure_mean_0, . . . , diastolic_blood_pressure_mean_23,

diastolic_blood_pressure_time_since_measured_0, . . . ,
diastolic_blood_pressure_time_since_measured_23]

- Eosinophils [eosinophils_mask_0, . . . , eosinophils_mask_23, eosinophils_mean_0, . . . , eosinophils_mean_23, eosinophils_time_since_measured_0, . . . , eosinophils_time_since_measured_23]
- Fibrinogen [fibrinogen_mask_0, . . . , fibrinogen_mask_23, fibrinogen_mean_0, . . . , fibrinogen_mean_23, fibrinogen_time_since_measured_0, . . . , fibrinogen_time_since_measured_23]
- Glucose [glucose_mask_0, . . . , glucose_mask_23, glucose_mean_0, . . . , glucose_mean_23, glucose_time_since_measured_0, . . . , glucose_time_since_measured_23]
- Hematocrit [hematocrit_mask_0, . . . , hematocrit_mask_23, hematocrit_mean_0, . . . , hematocrit_mean_23, hematocrit_time_since_measured_0, . . . , hematocrit_time_since_measured_23]
- Lymphocytes [lymphocytes_mask_0, . . . , lymphocytes_mask_23, lymphocytes_mean_0, . . . , lymphocytes_mean_23, lymphocytes_time_since_measured_0, . . . , lymphocytes_time_since_measured_23]
- Lymphocytes ascites [lymphocytes_ascites_mask_0, . . . , lymphocytes_ascites_mask_23, lymphocytes_ascites_mean_0, . . . , lymphocytes_ascites_mean_23, lymphocytes_ascites_time_since_measured_0, . . . , lymphocytes_ascites_time_since_measured_23]
- Atypical lymphocytes [lymphocytes_atypical_mask_0, . . . , lymphocytes_atypical_mask_23, lymphocytes_atypical_mean_0, . . . , lymphocytes_atypical_mean_23, lymphocytes_atypical_time_since_measured_0, . . . , lymphocytes_atypical_time_since_measured_23, lymphocytes_atypical_csl_mask_0, . . . , lymphocytes_atypical_csl_mask_23, lymphocytes_atypical_csl_ean_0, . . . , lymphocytes_atypical_csl_mean_23, lymphocytes_atypical_csl_time_since_measured_0, . . . , lymphocytes_atypical_csl_time_since_measured_23]
- Lymphocytes in body fluid [lymphocytes_body_fluid_mask_0, . . . , lymphocytes_body_fluid_mask_23, lymphocytes_body_fluid_mean_0, . . . , lymphocytes_body_fluid_mean_23, lymphocytes_body_fluid_time_since_measured_0, . . . , lymphocytes_body_fluid_time_since_measured_23]
- Lymphocytes percentage [lymphocytes_percent_mask_0, . . . , lymphocytes_percent_mask_23, lymphocytes_percent_mean_0, . . . , lymphocytes_percent_mean_23, lymphocytes_percent_time_since_measured_0, . . . , lymphocytes_percent_time_since_measured_23]
- Lymphocytes pleural [lymphocytes_pleural_mask_0, . . . , lymphocytes_pleural_mask_23, lymphocytes_pleural_mean_0, . . . , lymphocytes_pleural_mean_23, lymphocytes_pleural_time_since_measured_0, . . . , lymphocytes_pleural_time_since_measured_23]
- Magnesium [magnesium_mask_0, . . . , magnesium_mask_23, magnesium_mean_0, . . . , magnesium_mean_23, magnesium_time_since_measured_0, . . . , magnesium_time_since_measured_23]
- Mean corpuscular hemoglobin [mean_corpuscular_hemoglobin_mask_0, . . . , mean_corpuscular_hemoglobin_mask_23, mean_corpuscular_hemoglobin_mean_0, . . . , mean_corpuscular_hemoglobin_mean_23, mean_corpuscular_hemoglobin_time_since_measured_0, . . . , mean_corpuscular_hemoglobin_time_since_measured_23]
- Mean corpuscular hemoglobin concentration [mean_corpuscular_hemoglobin_concentration_mask_0, . . . , mean_corpuscular_hemoglobin_concentration_mask_23, mean_corpuscular_hemoglobin_concentration_mean_0, . . . , mean_corpuscular_hemoglobin_concentration_mean_23, mean_corpuscular_hemoglobin_concentration_time_since_measured_0, . . . , mean_corpuscular_hemoglobin_concentration_time_since_measured_23]

- Mean corpuscular volume [mean_corpuscular_volume_mask_0, . . . , mean_corpuscular_volume_mask_23, mean_corpuscular_volume_mean_0, . . . , mean_corpuscular_volume_mean_23, mean_corpuscular_volume_time_since_measured_0, . . . , mean_corpuscular_volume_time_since_measured_23]
- Monocytes [monocytes_mask_0, . . . , monocytes_mask_23, monocytes_mean_0, . . . , monocytes_mean_23, monocytes_time_since_measured_0, . . . , monocytes_time_since_measured_23, monocytes_csl_mask_0, . . . , monocytes_csl_mask_23, monocytes_csl_mean_0, . . . , monocytes_csl_mean_23, monocytes_csl_time_since_measured_0, . . . , monocytes_csl_time_since_measured_23]
- Neutrophils [neutrophils_mask_0, . . . , neutrophils_mask_23, neutrophils_mean_0, . . . , neutrophils_mean_23, neutrophils_time_since_measured_0, . . . , neutrophils_time_since_measured_23]
- Partial pressure of carbon dioxide [partial_pressure_of_carbon_dioxide_mask_0, . . . , partial_pressure_of_carbon_dioxide_mask_23, partial_pressure_of_carbon_dioxide_mean_0, . . . , partial_pressure_of_carbon_dioxide_mean_23, partial_pressure_of_carbon_dioxide_time_since_measured_0, . . . , partial_pressure_of_carbon_dioxide_time_since_measured_23]
- Partial thromboplastin [partial_thromboplastin_mask_0, . . . , partial_thromboplastin_mask_23, partial_thromboplastin_mean_0, . . . , partial_thromboplastin_mean_23, partial_thromboplastin_time_since_measured_0, . . . , partial_thromboplastin_time_since_measured_23]
- Peak inspiratory pressure [peak_inspiratory_pressure_mask_0, . . . , peak_inspiratory_pressure_mask_23, peak_inspiratory_pressure_mean_0, . . . , peak_inspiratory_pressure_mean_23, peak_inspiratory_pressure_time_since_measured_0, . . . , peak_inspiratory_pressure_time_since_measured_23]
- Ph in urine [ph_urine_mask_0, . . . , ph_urine_mask_23, ph_urine_mean_0, . . . , ph_urine_mean_23, ph_urine_time_since_measured_0, . . . , ph_urine_time_since_measured_23]
- Phosphate [phosphate_mask_0, . . . , phosphate_mask_23, phosphate_mean_0, . . . , phosphate_mean_23, phosphate_time_since_measured_0, . . . , phosphate_time_since_measured_23]
- Phosphorous [phosphorous_mask_0, . . . , phosphorous_mask_23, phosphorous_mean_0, . . . , phosphorous_mean_23, phosphorous_time_since_measured_0, . . . , phosphorous_time_since_measured_23]
- Plateau pressure [plateau_pressure_mask_0, . . . , plateau_pressure_mask_23, plateau_pressure_mean_0, . . . , plateau_pressure_mean_23, plateau_pressure_time_since_measured_0, . . . , plateau_pressure_time_since_measured_23]
- Platelets [platelets_mask_0, . . . , platelets_mask_23, platelets_mean_0, . . . , platelets_mean_23, platelets_time_since_measured_0, . . . , platelets_time_since_measured_23]
- Positive end expiratory pressure [positive_end_expiratory_pressure_mask_0, . . . , positive_end_expiratory_pressure_mask_23, positive_end_expiratory_pressure_mean_0, . . . , positive_end_expiratory_pressure_mean_23, positive_end_expiratory_pressure_time_since_measured_0, . . . , positive_end_expiratory_pressure_time_since_measured_23, positive_end_expiratory_pressure_set_mask_0, . . . , positive_end_expiratory_pressure_set_mask_23, positive_end_expiratory_pressure_set_mean_0, . . . , positive_end_expiratory_pressure_set_mean_23, positive_end_expiratory_pressure_set_time_since_measured_0, . . . , positive_end_expiratory_pressure_set_time_since_measured_23]

- Post void residual [post_void_residual_mask_0, . . . , post_void_residual_mask_23, post_void_residual_mean_0, . . . , post_void_residual_mean_23, post_void_residual_time_since_measured_0, . . . , post_void_residual_time_since_measured_23]
- Potassium [potassium_mask_0, . . . , potassium_mask_23, potassium_mean_0, . . . , potassium_mean_23, potassium_time_since_measured_0, . . . , potassium_time_since_measured_23]
- Potassium serum [potassium_serum_mask_0, . . . , potassium_serum_mask_23, potassium_serum_mean_0, . . . , potassium_serum_mean_23, potassium_serum_time_since_measured_0, . . . , potassium_serum_time_since_measured_23]
- Prothrombin time tested with INR [prothrombin_time_inr_mask_0, . . . , prothrombin_time_inr_mask_23, prothrombin_time_inr_mean_0, . . . , prothrombin_time_inr_mean_23, prothrombin_time_inr_time_since_measured_0, . . . , prothrombin_time_inr_time_since_measured_23]
- Prothrombin time using PT [prothrombin_time_pt_mask_0, . . . , prothrombin_time_pt_mask_23, prothrombin_time_pt_mean_0, . . . , prothrombin_time_pt_mean_23, prothrombin_time_pt_time_since_measured_0, . . . , prothrombin_time_pt_time_since_measured_23]
- Pulmonary artery pressure [pulmonary_artery_pressure_mask_0, . . . , pulmonary_artery_pressure_mask_23, pulmonary_artery_pressure_mean_0, . . . , pulmonary_artery_pressure_mean_23, pulmonary_artery_pressure_time_since_measured_0, . . . , pulmonary_artery_pressure_time_since_measured_23]
- Systolic pulmonary artery pressure[pulmonary_artery_pressure_systolic_mask_0, . . . , pulmonary_artery_pressure_systolic_mask_23, pulmonary_artery_pressure_systolic_mean_0, . . . , pulmonary_artery_pressure_systolic_mean_23, pulmonary_artery_pressure_systolic_time_since_measured_0, . . . , pulmonary_artery_pressure_systolic_time_since_measured_23]
- Pulmonary capillary wedge pressure [pulmonary_capillary_wedge_pressure_mask_0, . . . , pulmonary_capillary_wedge_pressure_mask_23, pulmonary_capillary_wedge_pressure_mean_0, . . . , pulmonary_capillary_wedge_pressure_mean_23, pulmonary_capillary_wedge_pressure_time_since_measured_0, . . . , pulmonary_capillary_wedge_pressure_time_since_measured_23]
- Red blood cell count [red_blood_cell_count_mask_0, . . . , red_blood_cell_count_mask_23, red_blood_cell_count_mean_0, . . . , red_blood_cell_count_mean_23, red_blood_cell_count_time_since_measured_0, . . . , red_blood_cell_count_time_since_measured_23]
- Red blood cell count ascites [red_blood_cell_count_ascites_mask_0, . . . , red_blood_cell_count_ascites_mask_23, red_blood_cell_count_ascites_mean_0, . . . , red_blood_cell_count_ascites_mean_23, red_blood_cell_count_ascites_time_since_measured_0, . . . , red_blood_cell_count_ascites_time_since_measured_23]
- Red blood cell count csf [red_blood_cell_count_csf_mask_0, . . . , red_blood_cell_count_csf_mask_23, red_blood_cell_count_csf_mean_0, . . . , red_blood_cell_count_csf_mean_23, red_blood_cell_count_csf_time_since_measured_0, . . . , red_blood_cell_count_csf_time_since_measured_23]
- Red blood cell count pleural [red_blood_cell_count_pleural_mask_0, . . . , red_blood_cell_count_pleural_mask_23, red_blood_cell_count_pleural_mean_0, . . . , red_blood_cell_count_pleural_mean_23, red_blood_cell_count_pleural_time_since_measured_0, . . . , red_blood_cell_count_pleural_time_since_measured_23]
- Red blood cell count in urine [red_blood_cell_count_urine_mask_0, . . . , red_blood_cell_count_urine_mask_23, red_blood_cell_count_urine_mean_0, . . . ,

red_blood_cell_count_urine_mean_23, red_blood_cell_count_urine_time_since_measured_0, . . . , red_blood_cell_count_urine_time_since_measured_23]

- Systemic vascular resistance [systemic_vascular_resistance_mask_0, . . . , systemic_vascular_resistance_mask_23, systemic_vascular_resistance_mean_0, . . . , systemic_vascular_resistance_mean_23, systemic_vascular_resistance_time_since_measured_0, . . . , systemic_vascular_resistance_time_since_measured_23]
- Tidal_volume_observed [tidal_volume_observed_mask_0, . . . , tidal_volume_observed_mask_23, tidal_volume_observed_mean_0, . . . , tidal_volume_observed_mean_23, tidal_volume_observed_time_since_measured_0, . . . , tidal_volume_observed_time_since_measured_23]
- Tidal volume [tidal_volume_set_mask_0, . . . , tidal_volume_set_mask_23, tidal_volume_set_mean_0, . . . , tidal_volume_set_mean_23, tidal_volume_set_time_since_measured_0, . . . , tidal_volume_set_time_since_measured_23]
- Tidal volume spontaneous [tidal_volume_spontaneous_mask_0, . . . , tidal_volume_spontaneous_mask_23, tidal_volume_spontaneous_mean_0, . . . , tidal_volume_spontaneous_mean_23, tidal_volume_spontaneous_time_since_measured_0, . . . , tidal_volume_spontaneous_time_since_measured_23]
- Total protein [total_protein_mask_0, . . . , total_protein_mask_23, total_protein_mean_0, . . . , total_protein_mean_23, total_protein_time_since_measured_0, . . . , total_protein_time_since_measured_23]
- Total protein in urine [total_protein_urine_mask_0, . . . , total_protein_urine_mask_23, total_protein_urine_mean_0, . . . , total_protein_urine_mean_23, total_protein_urine_time_since_measured_0, . . . , total_protein_urine_time_since_measured_23]
- Troponin_i [troponin_i_mask_0, . . . , troponin_i_mask_23, troponin_i_mean_0, . . . , troponin_i_mean_23, troponin_i_time_since_measured_0, . . . , troponin_i_time_since_measured_23]
- Troponin_t [troponin_t_mask_0, . . . , troponin_t_mask_23, troponin_t_mean_0, . . . , troponin_t_mean_23, troponin_t_time_since_measured_0, . . . , troponin_t_time_since_measured_23]
- Venous pvo2 [venous_pvo2_mask_0, . . . , venous_pvo2_mask_23, venous_pvo2_mean_0, . . . , venous_pvo2_mean_23, venous_pvo2_time_since_measured_0, . . . , venous_pvo2_time_since_measured_23]
- White blood cell count in urine [white_blood_cell_count_urine_mask_0, . . . , white_blood_cell_count_urine_mask_23, white_blood_cell_count_urine_mean_0, . . . , white_blood_cell_count_urine_mean_23, white_blood_cell_count_urine_time_since_measured_0, . . . , white_blood_cell_count_urine_time_since_measured_23]

E.6.2 Hospital Mortality

The selection procedure for the task 'Hospital Mortality' is discussed in detail in Appendix A.

Target: Hospital morality (that the patient dies at any point during this visit, even if they are discharged from the ICU to another unit in the hospital). [mort_hosp]

Shift: Insurance type (Medicare, Private, Medicaid, Government, Self Pay) [insurance]

List of causal features:

- Age in years [age]
- Gender [gender]
- Ethnicity [ethnicity]
- Height [height_mean_0]
- Weight [weight_mean_0]

List of arguably causal features:

- Bicarbonate [bicarbonate_mask_0..., bicarbonate_mask_23, bicarbonate_mean_0,...,bicarbonate_mean_23, bicarbonate_time_since_measured_0,..., bicarbonate_time_since_measured_23]
- Co2 [co2_mask_0..., co2_mask_23, co2_mean_0,...,co2_mean_23, co2_time_since_measured_0,..., co2_time_since_measured_23]
- Partial pressure of carbon dioxide (pCO2) and end_tidal CO2 (ETCO2) [co2_(etco2_pco2_etc)_mask_0..., co2_(etco2_pco2_etc)_mask_23, co2_(etco2_pco2_etc)_mean_0,..., co2_(etco2_pco2_etc)_mean_23, co2_(etco2_pco2_etc)_time_since_measured_0,..., co2_(etco2_pco2_etc)_time_since_measured_23]
- Partial pressure of oxygen [partial_pressure_of_oxygen_mask_0..., partial_pressure_of_oxygen_mask_23, partial_pressure_of_oxygen_mean_0,...,partial_pressure_of_oxygen_mean_23, partial_pressure_of_oxygen_time_since_measured_0,..., partial_pressure_of_oxygen_time_since_measured_23]
- Fraction inspired oxygen [fraction_inspired_oxygen_mask_0..., fraction_inspired_oxygen_mask_23, fraction_inspired_oxygen_mean_0,...,fraction_inspired_oxygen_mean_23, fraction_inspired_oxygen_time_since_measured_0,..., fraction_inspired_oxygen_time_since_measured_23, fraction_inspired_oxygen_set_mask_0..., fraction_inspired_oxygen_set_mask_23, fraction_inspired_oxygen_set_mean_0,...,fraction_inspired_oxygen_set_mean_23, fraction_inspired_oxygen_set_time_since_measured_0,..., fraction_inspired_oxygen_set_time_since_measured_23]
- Glascow coma score [glascow_coma_scale_total_mask_0..., glascow_coma_scale_total_mask_23, glascow_coma_scale_total_mean_0,...,glascow_coma_scale_total_mean_23, glascow_coma_scale_total_time_since_measured_0,..., glascow_coma_scale_total_time_since_measured_23]
- Lactate [lactate_mask_0..., lactate_mask_23, lactate_mean_0,...,lactate_mean_23, lactate_time_since_measured_0,..., lactate_time_since_measured_23]
- Lactic acid [lactic_acid_mask_0..., lactic_acid_mask_23, lactic_acid_mean_0,...,lactic_acid_mean_23, lactic_acid_time_since_measured_0,..., lactic_acid_time_since_measured_23]
- Sodium [sodium_mask_0..., sodium_mask_23, sodium_mean_0,...,sodium_mean_23, sodium_time_since_measured_0,..., sodium_time_since_measured_23]
- Hemoglobin [hemoglobin_mask_0..., hemoglobin_mask_23, hemoglobin_mean_0,...,hemoglobin_mean_23, hemoglobin_time_since_measured_0,..., hemoglobin_time_since_measured_23]
- Mean blood pressure [mean_blood_pressure_mask_0..., mean_blood_pressure_mask_23, mean_blood_pressure_mean_0,...,mean_blood_pressure_mean_23, mean_blood_pressure_time_since_measured_0,..., mean_blood_pressure_time_since_measured_23]
- Oxygen saturation [oxygen_saturation_mask_0..., oxygen_saturation_mask_23, oxygen_saturation_mean_0,...,oxygen_saturation_mean_23, oxygen_saturation_time_since_measured_0,..., oxygen_saturation_time_since_measured_23]
- Ph [ph_mask_0..., ph_mask_23, ph_mean_0,...,ph_mean_23, ph_time_since_measured_0,..., ph_time_since_measured_23]
- Respiratory rate [respiratory_rate_mask_0..., respiratory_rate_mask_23, respiratory_rate_mean_0,...,respiratory_rate_mean_23, respiratory_rate_time_since_measured_0,..., respiratory_rate_time_since_measured_23, respiratory_rate_set_mask_0..., respiratory_rate_set_mask_23, respiratory_rate_set_mean_0,...,respiratory_rate_set_mean_23,

respiratory_rate_set_time_since_measured_0,. . . ,
respiratory_rate_set_time_since_measured_23]

- Systolic blood pressure [systolic_blood_pressure_mask_0.. . . ,
systolic_blood_pressure_mask_23,
systolic_blood_pressure_mean_0,. . . ,systolic_blood_pressure_mean_23,
systolic_blood_pressure_time_since_measured_0,. . . ,
systolic_blood_pressure_time_since_measured_23]
- Heart rate [heart_rate_mask_0.. . . , heart_rate_mask_23,
heart_rate_mean_0,. . . ,heart_rate_mean_23, heart_rate_time_since_measured_0,. . . ,
heart_rate_time_since_measured_23]
- Temperature [temperature_mask_0.. . . , temperature_mask_23,
temperature_mean_0,. . . ,temperature_mean_23,
temperature_time_since_measured_0,. . . , temperature_time_since_measured_23]
- White blood cell count[white_blood_cell_count_mask_0.. . . ,
white_blood_cell_count_mask_23,
white_blood_cell_count_mean_0,. . . ,white_blood_cell_count_mean_23,
white_blood_cell_count_time_since_measured_0,. . . ,
white_blood_cell_count_time_since_measured_23]

List of other features:

- Height [height_mask_0.. . . , height_mask_23,
height_mean_1,. . . ,height_mean_23, height_time_since_measured_0,. . . ,
height_time_since_measured_23]
- Weight [weight_mask_0.. . . , weight_mask_23, weight_mean_1,. . . ,weight_mean_23,
weight_time_since_measured_0,. . . , weight_time_since_measured_23]
- Alanine aminotransferase [alanine_aminotransferase_mask_0.. . . ,
alanine_aminotransferase_mask_23,
alanine_aminotransferase_mean_0,. . . ,alanine_aminotransferase_mean_23,
alanine_aminotransferase_time_since_measured_0,. . . ,
alanine_aminotransferase_time_since_measured_23]
- Albumin [albumin_mask_0.. . . , albumin_mask_23,
albumin_mean_0,. . . ,albumin_mean_23, albumin_time_since_measured_0,. . . ,
albumin_time_since_measured_23]
- Alanine aminotransferase [albumin_ascites_mask_0.. . . , albumin_ascites_mask_23,
albumin_ascites_mean_0,. . . ,albumin_ascites_mean_23,
albumin_ascites_time_since_measured_0,. . . ,
albumin_ascites_time_since_measured_23]
- Albumin pleural [albumin_pleural_mask_0.. . . , albumin_pleural_mask_23,
albumin_pleural_mean_0,. . . ,albumin_pleural_mean_23,
albumin_pleural_time_since_measured_0,. . . ,
albumin_pleural_time_since_measured_23]
- Albumin in urine [albumin_urine_mask_0.. . . , albumin_urine_mask_23,
albumin_urine_mean_0,. . . ,albumin_urine_mean_23,
albumin_urine_time_since_measured_0,. . . , albumin_urine_time_since_measured_23]
- Alkaline phosphate [alkaline_phosphate_mask_0.. . . , alkaline_phosphate_mask_23,
alkaline_phosphate_mean_0,. . . ,alkaline_phosphate_mean_23,
alkaline_phosphate_time_since_measured_0,. . . ,
alkaline_phosphate_time_since_measured_23]
- Anion gap [anion_gap_mask_0.. . . , anion_gap_mask_23,
anion_gap_mean_0,. . . ,anion_gap_mean_23,
anion_gap_time_since_measured_0,. . . , anion_gap_time_since_measured_23]
- Asparate aminotransferase [asparate_aminotransferase_mask_0.. . . ,
asparate_aminotransferase_mask_23,
asparate_aminotransferase_mean_0,. . . ,asparate_aminotransferase_mean_23,
asparate_aminotransferase_time_since_measured_0,. . . ,
asparate_aminotransferase_time_since_measured_23]
- Basophils [basophils_mask_0.. . . , basophils_mask_23,
basophils_mean_0,. . . ,basophils_mean_23, basophils_time_since_measured_0,. . . ,
basophils_time_since_measured_23]

- Bilirubin [bilirubin_mask_0..., bilirubin_mask_23, bilirubin_mean_0,...,bilirubin_mean_23, bilirubin_time_since_measured_0,..., bilirubin_time_since_measured_23]
- Blood urea nitrogen [blood_urea_nitrogen_mask_0..., blood_urea_nitrogen_mask_23, blood_urea_nitrogen_mean_0,...,blood_urea_nitrogen_mean_23, blood_urea_nitrogen_time_since_measured_0,..., blood_urea_nitrogen_time_since_measured_23]
- Calcium [calcium_mask_0..., calcium_mask_23, calcium_mean_0,...,calcium_mean_23, calcium_time_since_measured_0,..., calcium_time_since_measured_23]
- Calcium ionized [calcium_ionized_mask_0..., calcium_ionized_mask_23, calcium_ionized_mean_0,...,calcium_ionized_mean_23, calcium_ionized_time_since_measured_0,..., calcium_ionized_time_since_measured_23]
- Calcium in urine [calcium_urine_mask_0..., calcium_urine_mask_23, calcium_urine_mean_0,...,calcium_urine_mean_23, calcium_urine_time_since_measured_0,..., calcium_urine_time_since_measured_23]
- Cardiac index [cardiac_index_mask_0..., cardiac_index_mask_23, cardiac_index_mean_0,...,cardiac_index_mean_23, cardiac_index_time_since_measured_0,..., cardiac_index_time_since_measured_23]
- Cardiac output by Fick principle [cardiac_output_fick_mask_0..., cardiac_output_fick_mask_23, cardiac_output_fick_mean_0,...,cardiac_output_fick_mean_23, cardiac_output_fick_time_since_measured_0,..., cardiac_output_fick_time_since_measured_23]
- Cardiac output by thermodilution [cardiac_output_thermodilution_mask_0..., cardiac_output_thermodilution_mask_23, cardiac_output_thermodilution_mean_0,...,cardiac_output_thermodilution_mean_23, cardiac_output_thermodilution_time_since_measured_0,..., cardiac_output_thermodilution_time_since_measured_23]
- Central venous pressure [central_venous_pressure_mask_0..., central_venous_pressure_mask_23, central_venous_pressure_mean_0,...,central_venous_pressure_mean_23, central_venous_pressure_time_since_measured_0,..., central_venous_pressure_time_since_measured_23]
- Chloride [chloride_mask_0..., chloride_mask_23, chloride_mean_0,...,chloride_mean_23, chloride_time_since_measured_0,..., chloride_time_since_measured_23]
- Chloride in urine [chloride_urine_mask_0..., chloride_urine_mask_23, chloride_urine_mean_0,...,chloride_urine_mean_23, chloride_urine_time_since_measured_0,..., chloride_urine_time_since_measured_23]
- Cholesterol [cholesterol_mask_0..., cholesterol_mask_23, cholesterol_mean_0,...,cholesterol_mean_23, cholesterol_time_since_measured_0,..., cholesterol_time_since_measured_23]
- HDL cholesterol [cholesterol_hdl_mask_0..., cholesterol_hdl_mask_23, cholesterol_hdl_mean_0,...,cholesterol_hdl_mean_23, cholesterol_hdl_time_since_measured_0,..., cholesterol_hdl_time_since_measured_23]
- LDL cholesterol [cholesterol_ldl_mask_0..., cholesterol_ldl_mask_23, cholesterol_ldl_mean_0,...,cholesterol_ldl_mean_23, cholesterol_ldl_time_since_measured_0,..., cholesterol_ldl_time_since_measured_23]
- Creatinine [creatinine_mask_0..., creatinine_mask_23, creatinine_mean_0,...,creatinine_mean_23, creatinine_time_since_measured_0,..., creatinine_time_since_measured_23]

- Creatinine ascites [creatinine_ascites_mask_0..., creatinine_ascites_mask_23, creatinine_ascites_mean_0,...,creatinine_ascites_mean_23, creatinine_ascites_time_since_measured_0,..., creatinine_ascites_time_since_measured_23]
- Creatinine body fluid [creatinine_body_fluid_mask_0..., creatinine_body_fluid_mask_23, creatinine_body_fluid_mean_0,...,creatinine_body_fluid_mean_23, creatinine_body_fluid_time_since_measured_0,..., creatinine_body_fluid_time_since_measured_23]
- Creatinine pleural [creatinine_pleural_mask_0..., creatinine_pleural_mask_23, creatinine_pleural_mean_0,...,creatinine_pleural_mean_23, creatinine_pleural_time_since_measured_0,..., creatinine_pleural_time_since_measured_23]
- Creatinine in urine [creatinine_urine_mask_0..., creatinine_urine_mask_23, creatinine_urine_mean_0,...,creatinine_urine_mean_23, creatinine_urine_time_since_measured_0,..., creatinine_urine_time_since_measured_23]
- Diastolic blood pressure [diastolic_blood_pressure_mask_0..., diastolic_blood_pressure_mask_23, diastolic_blood_pressure_mean_0,...,diastolic_blood_pressure_mean_23, diastolic_blood_pressure_time_since_measured_0,..., diastolic_blood_pressure_time_since_measured_23]
- Eosinophils [eosinophils_mask_0..., eosinophils_mask_23, eosinophils_mean_0,...,eosinophils_mean_23, eosinophils_time_since_measured_0,..., eosinophils_time_since_measured_23]
- Fibrinogen [fibrinogen_mask_0..., fibrinogen_mask_23, fibrinogen_mean_0,...,fibrinogen_mean_23, fibrinogen_time_since_measured_0,..., fibrinogen_time_since_measured_23]
- Glucose [glucose_mask_0..., glucose_mask_23, glucose_mean_0,...,glucose_mean_23, glucose_time_since_measured_0,..., glucose_time_since_measured_23]
- Hematocrit [hematocrit_mask_0..., hematocrit_mask_23, hematocrit_mean_0,...,hematocrit_mean_23, hematocrit_time_since_measured_0,..., hematocrit_time_since_measured_23]
- Lymphocytes [lymphocytes_mask_0..., lymphocytes_mask_23, lymphocytes_mean_0,...,lymphocytes_mean_23, lymphocytes_time_since_measured_0,..., lymphocytes_time_since_measured_23]
- Lymphocytes ascites [lymphocytes_ascites_mask_0..., lymphocytes_ascites_mask_23, lymphocytes_ascites_mean_0,...,lymphocytes_ascites_mean_23, lymphocytes_ascites_time_since_measured_0,..., lymphocytes_ascites_time_since_measured_23]
- Atypical lymphocytes [lymphocytes_atypical_mask_0..., lymphocytes_atypical_mask_23, lymphocytes_atypical_mean_0,...,lymphocytes_atypical_mean_23, lymphocytes_atypical_time_since_measured_0,..., lymphocytes_atypical_time_since_measured_23, lymphocytes_atypical_csl_mask_0..., lymphocytes_atypical_csl_mask_23, lymphocytes_atypical_csl_ean_0,...,lymphocytes_atypical_csl_mean_23, lymphocytes_atypical_csl_time_since_measured_0,..., lymphocytes_atypical_csl_time_since_measured_23]
- Lymphocytes in body fluid [lymphocytes_body_fluid_mask_0..., lymphocytes_body_fluid_mask_23, lymphocytes_body_fluid_mean_0,...,lymphocytes_body_fluid_mean_23, lymphocytes_body_fluid_time_since_measured_0,..., lymphocytes_body_fluid_time_since_measured_23]
- Lymphocytes percentage [lymphocytes_percent_mask_0..., lymphocytes_percent_mask_23,

lymphocytes_percent_mean_0,. . . ,lymphocytes_percent_mean_23,
lymphocytes_percent_time_since_measured_0,. . . ,
lymphocytes_percent_time_since_measured_23]

- Lymphocytes pleural [lymphocytes_pleural_mask_0.. . . ,
lymphocytes_pleural_mask_23,
lymphocytes_pleural_mean_0,. . . ,lymphocytes_pleural_mean_23,
lymphocytes_pleural_time_since_measured_0,. . . ,
lymphocytes_pleural_time_since_measured_23]
- Magnesium [magnesium_mask_0.. . . , magnesium_mask_23,
magnesium_mean_0,. . . ,magnesium_mean_23,
magnesium_time_since_measured_0,. . . , magnesium_time_since_measured_23]
- Mean corpuscular hemoglobin [mean_corpuscular_hemoglobin_mask_0.. . . ,
mean_corpuscular_hemoglobin_mask_23,
mean_corpuscular_hemoglobin_mean_0,. . . ,mean_corpuscular_hemoglobin_mean_23,
mean_corpuscular_hemoglobin_time_since_measured_0,. . . ,
mean_corpuscular_hemoglobin_time_since_measured_23]
- Mean corpuscular hemoglobin concentration
[mean_corpuscular_hemoglobin_concentration_mask_0.. . . ,
mean_corpuscular_hemoglobin_concentration_mask_23,
mean_corpuscular_hemoglobin_concentration_mean_0,. . . ,mean_corpuscular_hemoglobin_concentration_mean_2
mean_corpuscular_hemoglobin_concentration_time_since_measured_0,. . . ,
mean_corpuscular_hemoglobin_concentration_time_since_measured_23]
- Mean corpuscular volume [mean_corpuscular_volume_mask_0.. . . ,
mean_corpuscular_volume_mask_23,
mean_corpuscular_volume_mean_0,. . . ,mean_corpuscular_volume_mean_23,
mean_corpuscular_volume_time_since_measured_0,. . . ,
mean_corpuscular_volume_time_since_measured_23]
- Monocytes [monocytes_mask_0.. . . , monocytes_mask_23,
monocytes_mean_0,. . . ,monocytes_mean_23, monocytes_time_since_measured_0,. . . ,
monocytes_time_since_measured_23, monocytes_csl_mask_0.. . . ,
monocytes_csl_mask_23, monocytes_csl_mean_0,. . . ,monocytes_csl_mean_23,
monocytes_csl_time_since_measured_0,. . . ,
monocytes_csl_time_since_measured_23]
- Neutrophils [neutrophils_mask_0.. . . , neutrophils_mask_23,
neutrophils_mean_0,. . . ,neutrophils_mean_23,
neutrophils_time_since_measured_0,. . . , neutrophils_time_since_measured_23]
- Partial pressure of carbon dioxide [partial_pressure_of_carbon_dioxide_mask_0.. . . ,
partial_pressure_of_carbon_dioxide_mask_23, par-
tial_pressure_of_carbon_dioxide_mean_0,. . . ,partial_pressure_of_carbon_dioxide_mean_23,
partial_pressure_of_carbon_dioxide_time_since_measured_0,. . . ,
partial_pressure_of_carbon_dioxide_time_since_measured_23]
- Partial thromboplastin [partial_thromboplastin_mask_0.. . . ,
partial_thromboplastin_mask_23,
partial_thromboplastin_mean_0,. . . ,partial_thromboplastin_mean_23,
partial_thromboplastin_time_since_measured_0,. . . ,
partial_thromboplastin_time_since_measured_23]
- Peak inspiratory pressure [peak_inspiratory_pressure_mask_0.. . . ,
peak_inspiratory_pressure_mask_23,
peak_inspiratory_pressure_mean_0,. . . ,peak_inspiratory_pressure_mean_23,
peak_inspiratory_pressure_time_since_measured_0,. . . ,
peak_inspiratory_pressure_time_since_measured_23]
- Ph in urine [ph_urine_mask_0.. . . , ph_urine_mask_23,
ph_urine_mean_0,. . . ,ph_urine_mean_23, ph_urine_time_since_measured_0,. . . ,
ph_urine_time_since_measured_23]
- Phosphate [phosphate_mask_0.. . . , phosphate_mask_23,
phosphate_mean_0,. . . ,phosphate_mean_23, phosphate_time_since_measured_0,. . . ,
phosphate_time_since_measured_23]

- Phosphorous [phosphorous_mask_0..., phosphorous_mask_23, phosphorous_mean_0,...,phosphorous_mean_23, phosphorous_time_since_measured_0,..., phosphorous_time_since_measured_23]
- Plateau pressure [plateau_pressure_mask_0..., plateau_pressure_mask_23, plateau_pressure_mean_0,...,plateau_pressure_mean_23, plateau_pressure_time_since_measured_0,..., plateau_pressure_time_since_measured_23]
- Platelets [platelets_mask_0..., platelets_mask_23, platelets_mean_0,...,platelets_mean_23, platelets_time_since_measured_0,..., platelets_time_since_measured_23]
- Positive end expiratory pressure [positive_end_expiratory_pressure_mask_0..., positive_end_expiratory_pressure_mask_23, positive_end_expiratory_pressure_mean_0,...,positive_end_expiratory_pressure_mean_23, positive_end_expiratory_pressure_time_since_measured_0,..., positive_end_expiratory_pressure_time_since_measured_23, positive_end_expiratory_pressure_set_mask_0..., positive_end_expiratory_pressure_set_mask_23, positive_end_expiratory_pressure_set_mean_0,...,positive_end_expiratory_pressure_set_mean_23, positive_end_expiratory_pressure_set_time_since_measured_0,..., positive_end_expiratory_pressure_set_time_since_measured_23]
- Post void residual [post_void_residual_mask_0..., post_void_residual_mask_23, post_void_residual_mean_0,...,post_void_residual_mean_23, post_void_residual_time_since_measured_0,..., post_void_residual_time_since_measured_23]
- Potassium [potassium_mask_0..., potassium_mask_23, potassium_mean_0,...,potassium_mean_23, potassium_time_since_measured_0,..., potassium_time_since_measured_23]
- Potassium serum [potassium_serum_mask_0..., potassium_serum_mask_23, potassium_serum_mean_0,...,potassium_serum_mean_23, potassium_serum_time_since_measured_0,..., potassium_serum_time_since_measured_23]
- Prothrombin time tested with INR [prothrombin_time_inr_mask_0..., prothrombin_time_inr_mask_23, prothrombin_time_inr_mean_0,...,prothrombin_time_inr_mean_23, prothrombin_time_inr_time_since_measured_0,..., prothrombin_time_inr_time_since_measured_23]
- Prothrombin time using PT [prothrombin_time_pt_mask_0..., prothrombin_time_pt_mask_23, prothrombin_time_pt_mean_0,...,prothrombin_time_pt_mean_23, prothrombin_time_pt_time_since_measured_0,..., prothrombin_time_pt_time_since_measured_23]
- Pulmonary artery pressure [pulmonary_artery_pressure_mask_0..., pulmonary_artery_pressure_mask_23, pulmonary_artery_pressure_mean_0,...,pulmonary_artery_pressure_mean_23, pulmonary_artery_pressure_time_since_measured_0,..., pulmonary_artery_pressure_time_since_measured_23]
- Systolic pulmonary artery pressure[pulmonary_artery_pressure_systolic_mask_0..., pulmonary_artery_pressure_systolic_mask_23, pulmonary_artery_pressure_systolic_mean_0,...,pulmonary_artery_pressure_systolic_mean_23, pulmonary_artery_pressure_systolic_time_since_measured_0,..., pulmonary_artery_pressure_systolic_time_since_measured_23]
- Pulmonary capillary wedge pressure [pulmonary_capillary_wedge_pressure_mask_0..., pulmonary_capillary_wedge_pressure_mask_23, pulmonary_capillary_wedge_pressure_mean_0,...,pulmonary_capillary_wedge_pressure_mean_23, pulmonary_capillary_wedge_pressure_time_since_measured_0,..., pulmonary_capillary_wedge_pressure_time_since_measured_23]

- Red blood cell count [red_blood_cell_count_mask_0..., red_blood_cell_count_mask_23, red_blood_cell_count_mean_0,...,red_blood_cell_count_mean_23, red_blood_cell_count_time_since_measured_0,..., red_blood_cell_count_time_since_measured_23]
- Red blood cell count ascites [red_blood_cell_count_ascites_mask_0..., red_blood_cell_count_ascites_mask_23, red_blood_cell_count_ascites_mean_0,...,red_blood_cell_count_ascites_mean_23, red_blood_cell_count_ascites_time_since_measured_0,..., red_blood_cell_count_ascites_time_since_measured_23]
- Red blood cell count csf [red_blood_cell_count_csf_mask_0..., red_blood_cell_count_csf_mask_23, red_blood_cell_count_csf_mean_0,...,red_blood_cell_count_csf_mean_23, red_blood_cell_count_csf_time_since_measured_0,..., red_blood_cell_count_csf_time_since_measured_23]
- Red blood cell count pleural [red_blood_cell_count_pleural_mask_0..., red_blood_cell_count_pleural_mask_23, red_blood_cell_count_pleural_mean_0,...,red_blood_cell_count_pleural_mean_23, red_blood_cell_count_pleural_time_since_measured_0,..., red_blood_cell_count_pleural_time_since_measured_23]
- Red blood cell count in urine [red_blood_cell_count_urine_mask_0..., red_blood_cell_count_urine_mask_23, red_blood_cell_count_urine_mean_0,...,red_blood_cell_count_urine_mean_23, red_blood_cell_count_urine_time_since_measured_0,..., red_blood_cell_count_urine_time_since_measured_23]
- Systemic vascular resistance [systemic_vascular_resistance_mask_0..., systemic_vascular_resistance_mask_23, systemic_vascular_resistance_mean_0,...,systemic_vascular_resistance_mean_23, systemic_vascular_resistance_time_since_measured_0,..., systemic_vascular_resistance_time_since_measured_23]
- Tidal_volume_observed [tidal_volume_observed_mask_0..., tidal_volume_observed_mask_23, tidal_volume_observed_mean_0,...,tidal_volume_observed_mean_23, tidal_volume_observed_time_since_measured_0,..., tidal_volume_observed_time_since_measured_23]
- Tidal volume [tidal_volume_set_mask_0..., tidal_volume_set_mask_23, tidal_volume_set_mean_0,...,tidal_volume_set_mean_23, tidal_volume_set_time_since_measured_0,..., tidal_volume_set_time_since_measured_23]
- Tidal volume spontaneous [tidal_volume_spontaneous_mask_0..., tidal_volume_spontaneous_mask_23, tidal_volume_spontaneous_mean_0,...,tidal_volume_spontaneous_mean_23, tidal_volume_spontaneous_time_since_measured_0,..., tidal_volume_spontaneous_time_since_measured_23]
- Total protein [total_protein_mask_0..., total_protein_mask_23, total_protein_mean_0,...,total_protein_mean_23, total_protein_time_since_measured_0,..., total_protein_time_since_measured_23]
- Total protein in urine [total_protein_urine_mask_0..., total_protein_urine_mask_23, total_protein_urine_mean_0,...,total_protein_urine_mean_23, total_protein_urine_time_since_measured_0,..., total_protein_urine_time_since_measured_23]
- Troponin_i [troponin_i_mask_0..., troponin_i_mask_23, troponin_i_mean_0,...,troponin_i_mean_23, troponin_i_time_since_measured_0,..., troponin_i_time_since_measured_23]
- Troponin_t [troponin_t_mask_0..., troponin_t_mask_23, troponin_t_mean_0,...,troponin_t_mean_23, troponin_t_time_since_measured_0,..., troponin_t_time_since_measured_23]

- Venous pvo2 [venous_pvo2_mask_0..., venous_pvo2_mask_23, venous_pvo2_mean_0,...,venous_pvo2_mean_23, venous_pvo2_time_since_measured_0,..., venous_pvo2_time_since_measured_23]
- White blood cell count in urine [white_blood_cell_count_urine_mask_0..., white_blood_cell_count_urine_mask_23, white_blood_cell_count_urine_mean_0,...,white_blood_cell_count_urine_mean_23, white_blood_cell_count_urine_time_since_measured_0,..., white_blood_cell_count_urine_time_since_measured_23]

E.7 TableShift: NHANES

We have one task based on the National Health and Nutrition Examination Survey (NHANES) [Centers for Disease Control and Prevention, 2017].[14]

E.7.1 Childhood Lead

Target: Blood lead (ug/dL) [LBXBPB]

Shift: Binary indicator for whether family PIR (poverty-income ratio) is ≤ 1.3. [INDFMPIRBelowCutoff]

List of causal features:
- Country of birth [DMDBORN4]
- Age in years [RIDAGEYR]
- Gender [RIAGENDR]
- Race and hispanic origin [RIDRETH_merged]
- Year of survey [nhanes_year]

List of arguably causal features:
- Highest grade or level of school completed or highest degree received [DMDEDUC2]

List of other features:
- Marital status [DMDMARTL]

E.8 TableShift: Physionet

We have one task based on the 2019 PhysioNet Challenge [Reyna et al., 2020, 2019].[15] The data is released by PhysioNet [Goldberger et al., 2000].

E.8.1 Sepsis

Target: For septic patients, SepsisLabel is 1 if $t \geq t\_sepsis - 6$ and 0 if $t < t\_sepsis - 6$. For non-septic patients, SepsisLabel is 0. [SepsisLabel]

Shift: ICU length of stay (hours since ICU admission) [ICULOS]

List of causal features:
- Age (years) [Age]
- Gender [Gender]
- Administrative identifier for ICU unit (MICU); false (0) or true (1) [Unit1]
- Administrative identifier for ICU unit (SICU); false (0) or true (1) [Unit2]
- Time between hospital and ICU admission (hours since ICU admission) [HospAdmTime]

List of arguably causal features[16]*:*
- Temperature (deg C) [Temp]
- Leukocyte count (count/L) [WBC]
- Fibrinogen concentration (mg/dL) [Fibrinogen]
- Platelet count (count/mL) [Platelets]
- Heart rate (in beats per minute) [HR]
- Pulse oximetry (%) [O2Sat]

[14] `https://www.cdc.gov/nchs/nhanes/index.htm`

[15] `https://physionet.org/content/challenge-2019/1.0.0/`

- Systolic BP (mm Hg) [SBP]
- Mean arterial pressure (mm Hg) [MAP]
- Diastolic BP (mm Hg) [DBP]
- Respiration rate (breaths per minute) [Resp]
- End tidal carbon dioxide (mm Hg) [EtCO2]
- Excess bicarbonate (mmol/L) [BaseExcess]
- Bicarbonate (mmol/L) [HCO3]
- Fraction of inspired oxygen (%) [FiO2]
- pH [pH]
- Partial pressure of carbon dioxide from arterial blood (mm Hg) [PaCO2]
- Oxygen saturation from arterial blood (%) [SaO2]
- Aspartate transaminase (IU/L) [AST]
- Blood urea nitrogen (mg/dL) [BUN]
- Alkaline phosphatase (IU/L) [Alkalinephos]
- Calcium (mg/dL) [Calcium]
- Chloride (mmol/L) [Chloride]
- Creatinine (mg/dL) [Creatinine]
- Direct bilirubin (mg/dL) [Bilirubin_direct]
- Serum glucose (mg/dL) [Glucose]
- Lactic acid (mg/dL) [Lactate]
- Magnesium (mmol/dL) [Magnesium]
- Phosphate (mg/dL) [Phosphate]
- Potassium (mmol/L) [Potassium]
- Total bilirubin (mg/dL) [Bilirubin_total]
- Troponin I (ng/mL) [TroponinI]
- Hematocrit (
- Hemoglobin (g/dL) [Hgb]
- Partial thromboplastin time (seconds) [PTT]

List of other features:
- The training set (i.e. hospital) from which an example is drawn [set]

E.9 TableShift: UCI

We have one task based on a dataset by Strack et al. [2014] from the UCI Machine Learning Repository [Clore et al., 2014].[17]

E.9.1 Hospital Readmission

Target: No record of readmission [readmitted]

Shift: Admission source [admission_source_id]

List of causal features:
- Race [race]
- Gender [gender]
- Age [age]
- Payer code [payer_code]
- Medical specialty of the admitting physician [medical_specialty]

List of arguably causal features:
- Weight in pounds [weight]
- Primary diagnosis [diag_1]
- Secondary diagnosis [diag_2]
- Additional secondary diagnosis [diag_3]

[17] `https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008`

- Total number of diagnoses [number_diagnoses]
- Discharge type [discharge_disposition_id]
- Count of days between admission and discharge [time_in_hospital]
- Number of outpatient visits of the patient in the year preceding the encounter [number_outpatient]
- Number of emergency visits of the patient in the year preceding the encounter [number_emergency]
- Number of inpatient visits of the patient in the year preceding the encounter [number_inpatient]
- Max glucose serum [max_glu_serum]
- Hemoglobin A1c test result [A1Cresult]
- Change in metformin medication [metformin]
- Change in repaglinide medication [repaglinide]
- Change in nateglinide medication [nateglinide]
- Change in chlorpropamide medication [chlorpropamide]
- Change in glimepiride medication [glimepiride]
- Change in acetohexamide medication [acetohexamide]
- Change in glipizide medication [glipizide]
- Change in glyburide medication [glyburide]
- Change in tolbutamide medication [tolbutamide]
- Change in pioglitazone medication [pioglitazone]
- Change in rosiglitazone medication [rosiglitazone]
- Change in acarbose medication [acarbose]
- Change in miglitol medication [miglitol]
- Change in troglitazone medication [troglitazone]
- Change in tolazamide medication [tolazamide]
- Change in examide medication [examide]
- Change in citoglipton medication [citoglipton]
- Change in insulin medication [insulin]
- Change in glyburide_metformin medication [glyburide_metformin]
- Change in glipizide_metformin medication [glipizide_metformin]
- Change in glimepiride_pioglitazone medication [glimepiride_pioglitazone]
- Change in metformin_rosiglitazone medication [metformin_rosiglitazone]
- Change in metformin_pioglitazone medication [metformin_pioglitazone]
- Change in any medication [change]
- Diabetes medication prescribed [diabetesMed]

List of other features:
- Admission type [admission_type_id]
- Number of lab tests performed during the encounter [num_lab_procedures]
- Number of procedures (other than lab tests) performed during the encounter [num_procedures]
- Number of distinct generic drugs administered during the encounter [num_medications]

E.10 MEPS

We have one task based on the Medical Expenditure Panel Survey (MEPS) [Agency for Healthcare Research and Quality, 2019].

E.10.1 Utilization

Dataset. We consider the MEPS 2019 Full Year Consolidated Data File. The dataset contains information on individuals taking part in one of the two MEPS panels in 2019. In particular, these individuals belong either to Panel 23 in its 3-5 round, or to Panel 24 in its 1-3 round. We train on the

first round in 2019 for each panel, that is, Round 3 of Panel 23 and Round 1 of Panel 24, and predict the total health care utilization across the year 2019. We adapt the target definition by Hardt and Kim [2023].

Distribution shift. We split the domains by health insurance type, analogous to *TableShift* in the task 'Stay in ICU' and 'Hospital Mortality'. We train on individuals with public health insurance, and use individuals with private health insurance as testing domain.

Target: Measure of health care utilization > 3 [TOTEXP19]

Shift: Insurance type [INSCOV19]

List of causal features:
- Sex [SEX]
- Race [RACEV1X, RACEV2X, RACEAX, RACEBX, RACEWX, RACETHX]
- Hispanic ethnicity [HISPANX, HISPNCAT]
- Years of education [EDUCYR]
- Educational attainment [HIDEG]
- Paid sick leaves [SICPAY31]
- Paid leave to visit doctor [PAYDR31]
- Person is born in U.S. [BORNUSA]
- Years person lived in the U.S. [YRSINUS]
- How well person speaks English [HWELLSPK]
- Speak other language at home [OTHLGSPK]
- What language spoken other than English [WHTLGSPK]
- Region [REGION31]
- Age [AGE31X]

List of arguably causal features:
- Family size [FCSZ1231, FAMSZE31]
- Martial status [MARRY31X]
- Flexible Spending Accounts [FSAGT31, HASFSA31, PFSAMT31]
- Employer offers health insurance [OFREMP31, OFFER31X]
- Insurance coverage from current main job [CMJHLD31]
- Covered by Medicare [MCARE31, MCRPD31, MCRPB31, MCRPHO31, MCARE31X, MCRPD31X]
- Covered by Medicaid [MCAID31, MCDHMO31, MCDMC31, MCAID31X, MCDAT31X]
- Covered by TRICARE/CHAMPVA [TRIAT31X, TRICR31X, TRILI31X, TRIST31X, TRIST31X, TRIPR31X, TRIEX31X, TRICH31X]
- Detailed type of covering entity [PRVHMO31, GOVTA31, GOVAAT31, GOVTB31, GOVBAT31, GOVTC31, GOVCAT31, VAPROG31, VAPRAT31, IHS31, IHSAT31, PRIDK31, PRING31, PUB31X, PUBAT31X, PRIEU31, PRIOG31, PRSTX31, PRINEO31, PRIEUO31, PRIV31, PRIVAT31, DISVW31X,]
- Health insurance held from current main job [HELD31X]
- Insured [INS31X, INSAT31X]
- Dental insurance [DENTIN31, DENTIN31, DNTINS31]
- Prescription drug private insurance [PMEDIN31, PMDINS31, PMEDUP31, PMEDPY31]
- Pension Plan [RETPLN31]
- Employment status [EMPST31]
- Student status [FTSTU31X]
- Has more than one job [MORJOB31]
- Difference in wage by round [DIFFWG31]
- Updated hourly wage [NHRWG31]
- Hourly wage of current main job [HRWG31X]
- Hours per week [HOUR31]

- Temporary current main job [TEMPJB31]
- Seasonal current main job [SSNLJB31]
- Self-employed [SELFCM31]
- Choice of health plans [CHOIC31]
- Industry group [INDCAT31]
- Occupation group [OCCCAT31]
- Union status [UNION31]
- Reason for not working [NWK31]
- Paid vacation [PAYVAC31]
- Instrumental Activities of Daily Living (IADL) help [IADLHP31]
- Activities of Daily Living (ADL) help [ADLHLP31]
- Use of assistive technology [AIDHLP31]
- Limitations in physical functioning [WLKLIM31, LFTDIF31, STPDIF31, WLKDIF31, MILDIF31, BENDIF31, RCHDIF31, FNGRDF31, ACTLIM31]
- Social limitations [SOCLIM31]
- Work, housework, and school limitations [WRKLIM31, WRKLIM31, HSELIM31, SCHLIM31, UNABLE31]
- Cognitive limitations [COGLIM31]
- Priority condition variables [ASTHEP31, ASSTIL31, ASATAK31, CHBRON31]
- Asthma medications [ASMRCN31, ASPREV31, ASDALY31, ASPKFL31, ASEVFL31, ASWNFL31, ASACUT31]
- Active duty in military [ACTDTY31]
- Perceived health status [RTHLTH31]
- Perceived mental health status [MNHLTH31]

List of other features:

- Current main job at private for-profit, nonprofit, or a government entity [JOBORG31]
- Self-employed business is incorporated, a proprietorship, or a partnership [BSNTY31]
- Number of employees [NUMEMP31]
- Firm has more than one location [MORE31]
- Month started current main job [STJBMM31]
- Year started current main job [STJBYY31]
- Veterans Specific Activity Questionnaire (VASQ) [VACMPY31, VAPROX31, VASPUN31, VACMPM31, VASPMH31, VASPOU31, VAPRHT31, VAWAIT31, VAWAIT31, VALOCT31, VANTWK31, VANEED31, VAOUT31, VAPAST31, VACOMP31, VAMREC31, VAGTRC31, VACARC31, VAPROB31, VAREP31, VACARE31, VAPCPR31, VAPROV31, VAPCOT31, VAPCCO31, VAPCRC31, VAPCSN31, VAPCRF31, VAPCSO31, VAPCOU31, VAPCUN31, VASPCL31, VAPACT31, VACTDY31, VARECM31, VAMOBL31, VACOPD31, VADERM31, VAGERD31, VAHRLS31, VABACK31, VAJTPN31, VARTHR31, VAGOUT31, VANECK31, VAFIBR31, VATMD31, VACOST31, VAPTSD31, VABIPL31, VADEPR31, VAMOOD31, VAPROS31, VARHAB31, VAMNHC31, VAGCNS31, VARXMD31, VACRGV31, VALCOH31]
- Data collection round [RNDFLG31]
- Imputation flag [HRWGIM31]
- How hourly wage was calculated [HRHOW31]
- Verification [VERFLG31]
- Survey related information [REFPRS31, REFRL31X, FCRP1231, FMRS1231, FAMS1231, RESP31, PROXY31, BEGRFM31, BEGRFY31, ENDRFM31, ENDRFY31, INSCOP31, INSC1231, ELGRND31, MOPID31X, DAPID31X]
- Round [RUSIZE31, RUSIZE31, RUCLAS31, PSTATS31, SPOUID31, SPOUIN31]

E.11 SIPP

We have one task based on the Survey of Income and Program Participation (SIPP) [U.S. Census Bureau, 2014].

E.11.1 Poverty

Dataset. We work with Wave 1 and Wave 2 of the SIPP 2014 panel data. We train on Wave 1 and want to predict whether an individual has an official poverty measure larger than the median in Wave 2 [Hardt and Kim, 2023].

Distribution shift. We use individuals with U.S. citizenship as the training domain, and individuals without U.S. citizenship as testing domain. This simulates a survey collection with a biased sample, e.g. individuals without U.S. citizenship are systematically excluded.

Target: Household income-to-poverty ratio ≥ 3 [OPM_RATIO]

Shift: Citizenship status [CITIZENSHIP_STATUS]

List of causal features:
- Marital status [MARITAL_STATUS]
- Educational attainment [EDUCATION]
- Race[RACE]
- Gender [GENDER]
- Age [AGE]
- Spanish, Hispanic, or Latino[ORIGIN]
- Disability status [HEALTHDISAB]
- Hearing difficulties [HEALTH_HEARING]
- Vision difficulties [HEALTH_SEEING]
- Cognitive difficulties [HEALTH_COGNITIVE]
- Ambulatory difficulties [HEALTH_AMBULATORY]
- Difficulties in self-care [HEALTH_SELF_CARE]
- Difficulties in doing errands [HEALTH_ERRANDS_DIFFICULTY]
- Core disability [HEALTH_CORE_DISABILITY]
- Supplemental disability [HEALTH_SUPPLEMENTAL_DISABILITY]

List of arguably causal features:
- Household income [HOUSEHOLD_INC]
- Family size [FAMILY_SIZE_AVG]
- Received worker's compensation [RECEIVED_WORK_COMP]
- Unemployment compensation [UNEMPLOYMENT_COMP]
- Amount of unemployment compensation [UNEMPLOYMENT_COMP_AMOUNT]
- Severance pay and pension[SEVERANCE_PAY_PENSION]
- Amount for forster child care [FOSTER_CHILD_CARE_AMT]
- Amount for child support [CHILD_SUPPORT_AMT]
- Alimony amount [ALIMONY_AMT]
- Income [INCOME]
- Income from assistance [INCOME_FROM_ASSISTANCE]
- Amount of savings and investments [SAVINGS_INV_AMOUNT]
- Amount of veteran benefits [VA_BENEFITS_AMOUNT]
- Amount of retirement income [RETIREMENT_INCOME_AMOUNT]
- Amount of survivor income [SURVIVOR_INCOME_AMOUNT]
- Amount of disability benefits [DISABILITY_BENEFITS_AMOUNT]
- Percentage of year in which individual received assistance from MEDICARE [MEDICARE_ASSISTANCE]
- Number of sick days [DAYS_SICK]
- Number of hospital nights [HOSPITAL_NIGHTS]
- Number of presciptions for medicaments [PRESCRIPTION_MEDS]
- Number of dentist visits[VISIT_DENTIST_NUM]
- Number of doctor visits [VISIT_DOCTOR_NUM]
- Amount paid for non-premium medical out-of-pocket expenditures [HEALTH_OVER_THE_COUNTER_PRODUCTS_PAY]

- Amount paid medical care [HEALTH_MEDICAL_CARE_PAY]
- Amount paid for health insurance premiums [HEALTH_INSURANCE_PREMIUMS]
- Amount of social security benefits [SOCIAL_SEC_BENEFITS]
- Transportation assistance [TRANSPORTATION_ASSISTANCE]
- Own living quarters [LIVING_OWNERSHIP]

List of anti-causal features:
- Type of living quarters [LIVING_QUARTERS_TYPE]
- Percentage of year in which individual received assistance from TANF [TANF_ASSISTANCE]
- Percentage of year in which individual received food assistance[FOOD_ASSISTANCE]
- Percentage of year in which individual received assistance from SNAP [SNAP_ASSISTANCE]
- Percentage of year in which individual received assistance from WIC [WIC_ASSISTANCE]
- Percentage of year in which individual received assistance from MEDICAID [MEDICAID_ASSISTANCE]

E.12 Details on distribution shifts

We provide Table 6 with details on the observed distribution shifts. We adapt the metrics for target shift, concept shift and covariate shift from Gardner et al. [2023]. See Appendix E.2 of their paper for the detailed definitions. For selected tasks, we give additional insights into the concept shift by detailing it on the variable level.[18] See Figure 52 to Figure 55. We note that we conducted the in-depth analysis of the distribution shift *post* selecting the causal features and running our experiments described in Section 3 and Appendix C. Our code is based on an unpublished script by Gardner et al. [2023].

Table 6: Summary of tasks and their associated distribution shifts.

Task	Covariate shift (OTDD)	Concept shift (FDD)	Label shift (L2 distance)
Food Stamps	14.20	640.82	0.0008
Income	30.60	1.40	0.0060
Public Coverage	5.79	4.06	0.1701
Unemployment	75.47	13,389,512.51	0.0003
ANES	13.60	2.23	0.0025
Diabetes	12.28	0.10	0.0332
Hypertension	4.69	0.04	0.0022
Hospital Readmission	42.37	1.30	0.0060
Childhood Lead	1.30	0.01	0.0026
Sepsis	6609.73	8.44	0.0040
ICU Length of Stay	56,439,324,672.00	47,042,729,585.25	0.0033
ICU Hospital Mortality	64,479,092,736.00	42,639,188,407.47	0.0015
ASSISTments	24,054.59	1137.42	0.0670
College Scorecard	43,566.39	2116.63	0.0337
SIPP	6,344,306.0	5,752,406.89	0.0751
MEPS	66.28	4.01	0.0013

[18] We perform the analysis for tasks with less than 100 features due to computational costs

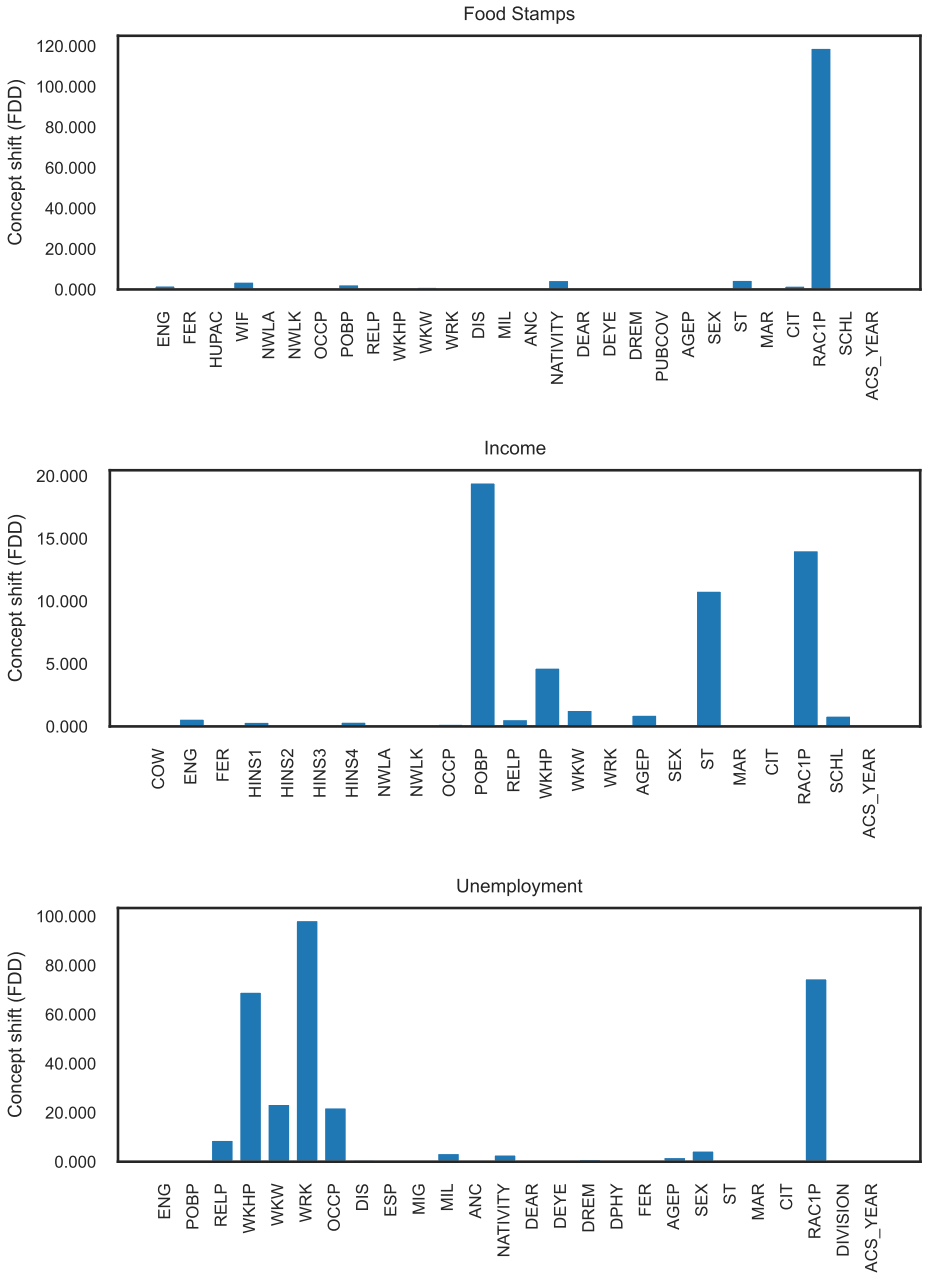


Figure 52: Concept shift on a variable level. Measured in FDD distance [Gardner et al., 2023].

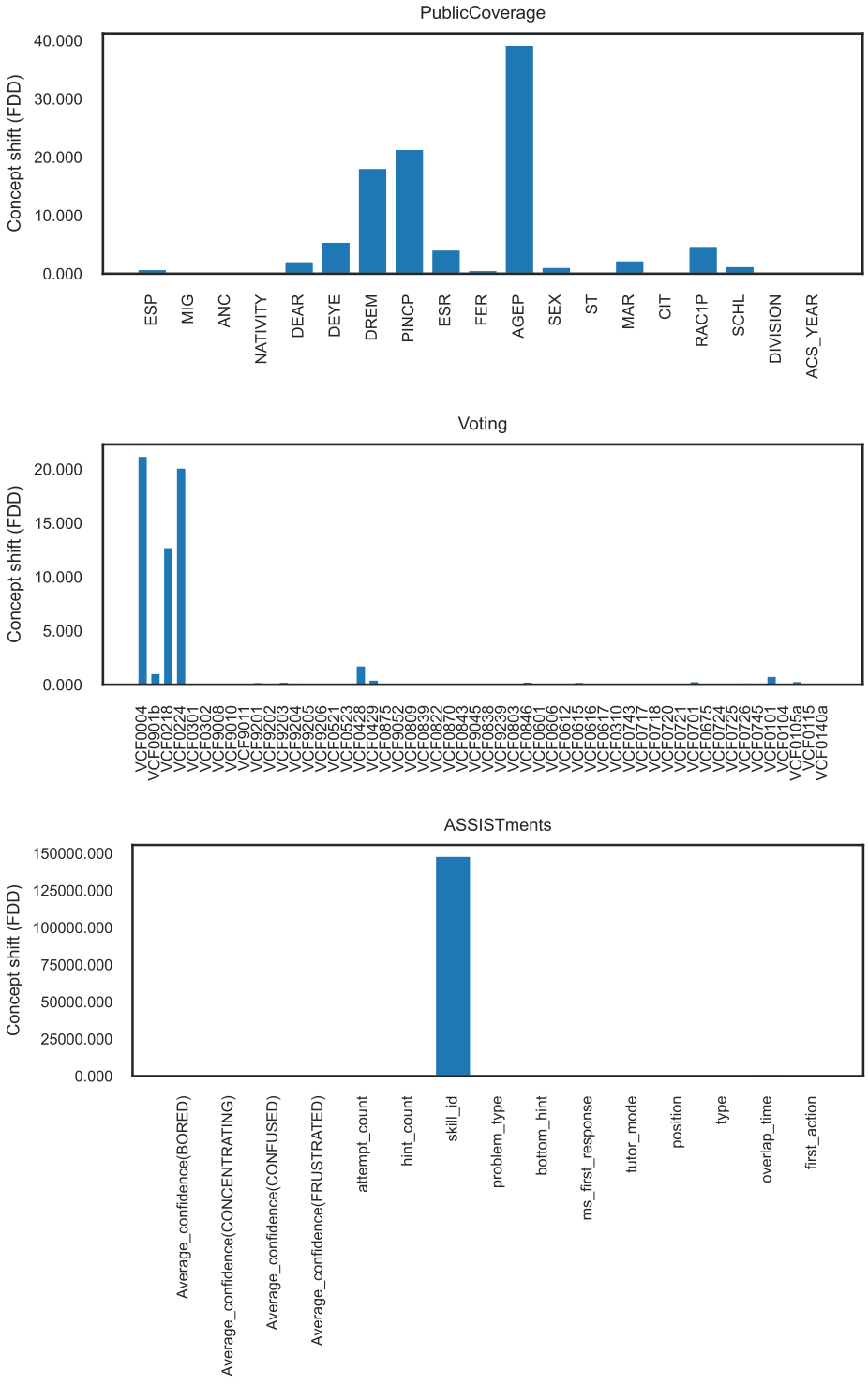


Figure 53: Concept shift on a variable level. Measured in FDD distance [Gardner et al., 2023]. (Continued)

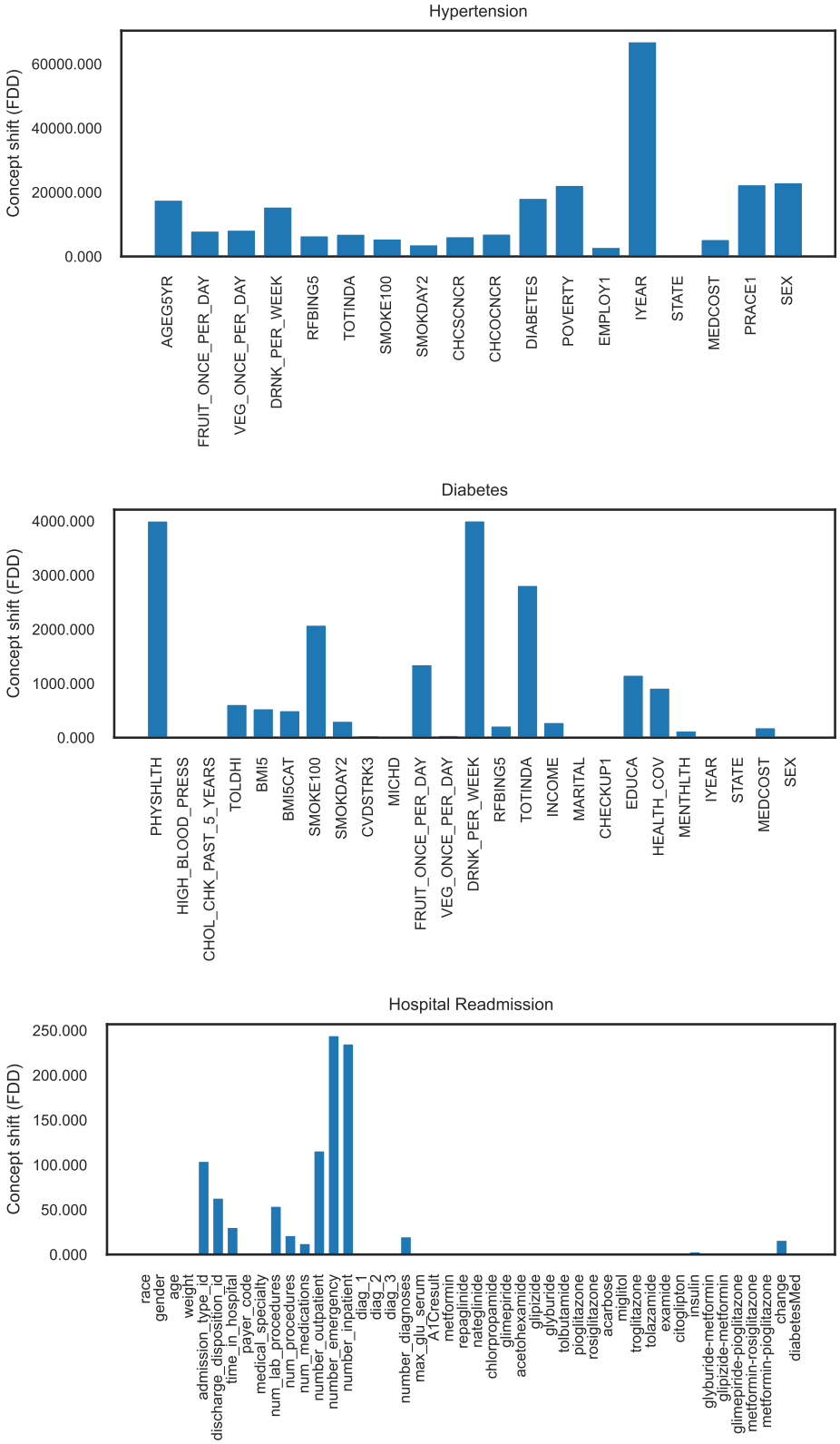


Figure 54: Concept shift on a variable level. Measured in FDD distance [Gardner et al., 2023]. (Continued)

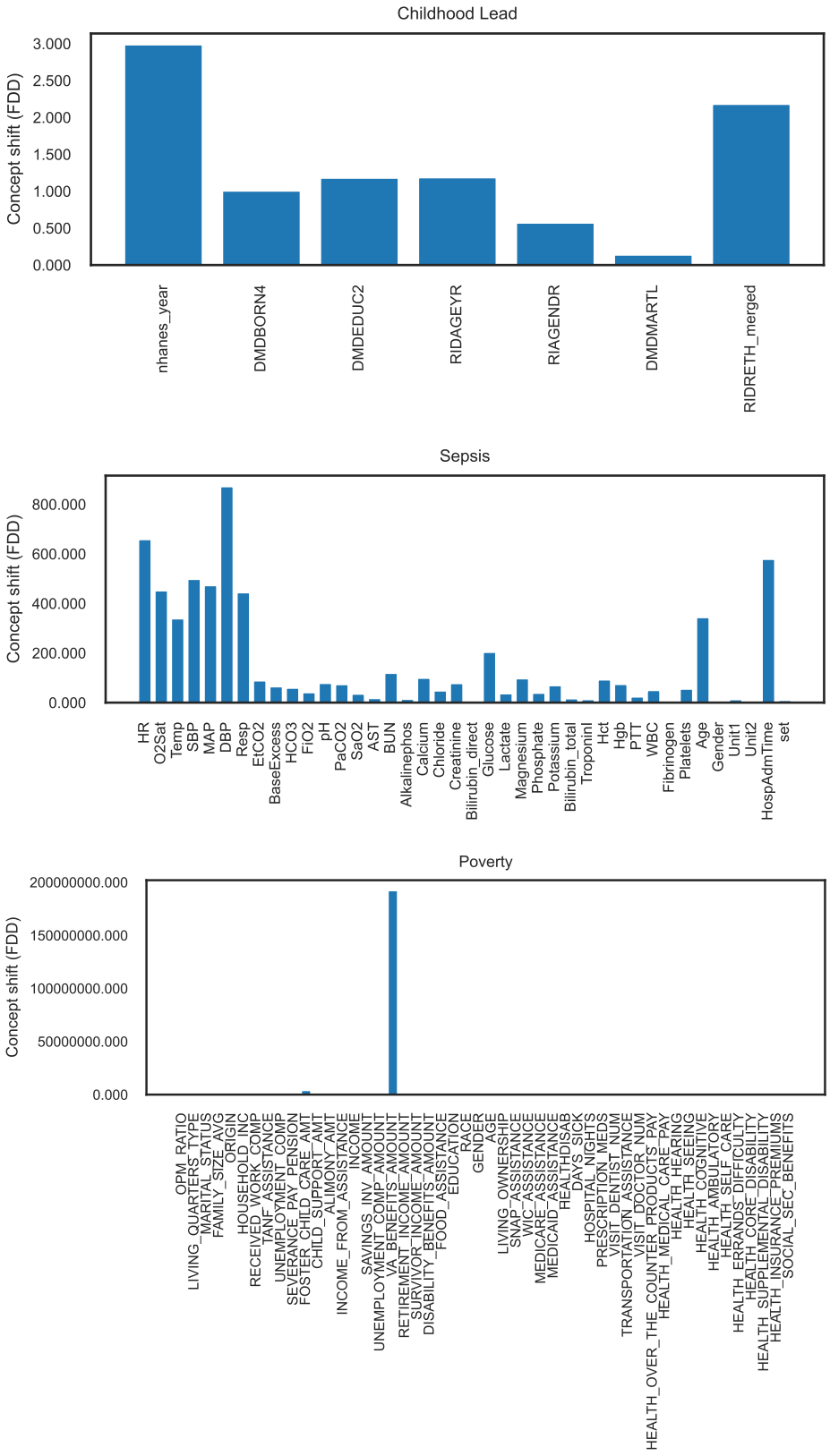


Figure 55: Concept shift on a variable level. Measured in FDD distance [Gardner et al., 2023]. (Continued)

Table 7: Details to data sources, access and licenses.

Task	Data Source	Data access	License
Food Stamps	American Community Survey	Public	CC0
Income	American Community Survey	Public	CC0
Public Coverage	American Community Survey	Public	CC0
Unemployment	American Community Survey	Public	CC0
Voting	American National Election Studies	Restricted-use	Unknown
Diabetes	Behavioral Risk Factor Surveillance System	Public	Open Data Commons Open Database License
Hypertension	Behavioral Risk Factor Surveillance System	Public	Open Data Commons Open Database License
College Scorecard	U.S. Department of Education	Public	Creative Commons Attribution License
ASSISTments	Kaggle	Public	Unknown
Stay in ICU	Medical Information Mart for Intensive Care	Restricted-use	PhysioNet Credentialed Health Data License
Hospital Mortality	Medical Information Mart for Intensive Care	Restricted-use	PhysioNet Credentialed Health Data License
Hospital Readmission	UCI Machine Learning Repository	Public	Creative Commons Attribution License
Childhood Lead	National Health and Nutrition Examination Survey	Restricted public	Open Database License
Sepsis	PhysioNet	Public	Creative Commons Attribution License
Utilization	Medical Expenditure Panel Survey	Restricted public	Open Data Commons Open Database License
Poverty	Survey of Income and Program Participation	Public	Unknown

Table 8: Description of tasks.

Task	Target	Shift	In-domain	Out-of-domain	Shift Gap	Obs.
Food Stamps	Food stamp recipiency in past year for households with child	Geographic region (U.S. divisions)	New England, Middle Atlantic, East North Central, West North Central, South Atlantic, West South Central, Mountain, Pacific	East South Central	2.90%	840,582
Income	Income $\geq$ 56k for employed adults	Geographic region (U.S. Divisions)	Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, Pacific	New England	7.71%	1,664,500
Public Health Insurance	Coverage of non-Medicare eligible low-income individuals	Disability status	Without a disability	With a disability	14.00%	5,916,565
Unemployment	Unemployment for non-social security-eligible adults	Education level	High school diploma or higher	No high school diploma	17.69%	
Voting	Voted in U.S. presidential election	Geographic region (U.S. regions)	Northeast, North Central, West	South	11.11%	8,280
Diabetes	Diabetes diagnosis	Race	White	Black or African American, American Indian or Alaskan Native, Asian, Native Hawaiian or other Pacific Islander, Other race	4.70%	1,444,176

Continued on next page

Table 8: Description of tasks. (Continued)

Task	Target	Shift	In-domain	Out-of-domain	Shift Gap	Obs.
Hypertension	Hypertension diagnosis for high-risk age (50+)	BMI category	Underweight, normal weight	Overweight, obese	1.37%	846,761
College Scorecard	Low degree completion rate	Carnegie classification	Different institution, e.g., special focus institutions (health professions), Master's colleges & universities (medium programs),	Special focus institutions (schools of art, music, and design, theological seminaries, bible colleges, and other faith-related institutions, others), Baccalaureate/Associate's colleges, Master's colleges & universities (larger programs)	18.36%	124,699
ASSISTments	Next answer correct	School	$\approx$ 700 schools	10 schools	13.18%	2,667,776
Stay in ICU	Length of stay $\geq$ 3 hrs in ICU	Insurance type	Private, Medicaid, Government, Self Pay	Medicare	5.78%	23,944
Hospital Mortality	ICU patient expires in hospital during current visit	Insurance type	Private, Medicaid, Government, Self Pay	Medicare	3.85%	23,944
Hospital Readmission	30-day readmission of diabetic hospital patients	Admission source	Different admission sources, e.g., physician referral, clinic referral, transfer from a hospital, court/law enforcement, transfer from hospice	Emergency room	7.77%	99,493
Childhood Lead	Blood lead levels above CDC blood level reference value	Poverty level	Poverty-income ratio > 1.3	Poverty-income ratio $\leq$ 1.3	4.82%	27,499

Continued on next page

Table 8: Description of tasks. (Continued)

Task	Target	Shift	In-domain	Out-of-domain	Shift Gap	Obs.
Sepsis	Sepsis onset within next 6hrs for hospital patients	Length of stay	Having been in ICU for $\leq$ 47 hours	Having been in ICU for > 47 hours	6.40%	1,552,210
Utilization	Measure of health care utilization > 3	Insurance type	Any public	Private only	-4.01%	28,512
Poverty	Household income-to-poverty ratio $\geq$ 3	Citizenship status	Citizen of the U.S.	Not citizen of the U.S.	21.59%	39,720

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide details in Section 3 to the claims made in abstract and introduction. We point out in the introduction that our study is based on domain-knowledge selection of causal features, and empirical observations might not hold for other datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of our study in the Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not have any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We explain the experimental procedure in Section 2.4, and in more detail in Appendix B. We give a list of our feature selections along with the datasets and distribution shifts in Appendix E. The code to replicate the experiments is provided at `https://github.com/socialfoundations/causal-features`. We link the reader to the datasets used in our study in Appendix E; most of them have public access.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to the code to reproduce our empirical results and instructions at `https://github.com/socialfoundations/causal-features`. We link the reader to the raw datasets in Appendix E. Most of them are public accessible (11/16 tasks), some allow only restricted use (5/11 tasks). We give details on steps needed for the preprocessing at `https://github.com/socialfoundations/causal-features/blob/main/README.md`. Experimental run details are provided in Appendix B.3.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 2.4 and more details in Appendix B. The code is provided at `https://github.com/socialfoundations/causal-features`.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report all results accompanied by (approximate) confidence intervals. We describe the confidence intervals briefly in Section 3, and provide details in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the computer resources in Appendix B.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research does not involve human subjects or participants. We do not create a dataset, but respect the terms of datasets used and link to their licenses.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Causal inference is an important tool in policymaking and as such it has large potential societal impact, both positive and negative. Our work helps inform the use of some causal methods by showing that these methods may not achieve better performance than standard predictors. At the same time, we emphasize in Section 4 that our analysis applies to questions of domain generalization for a particular range of settings we considered. We caution that our findings should not be applied too broadly to causal inference and should not deter the practitioner from engaging in rigorous causal inference.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work is foundational research and therefore, poses no high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the creators of models, evaluated in our paper, in Section 2.3 and 3. Creators of code, used in our paper, are credited in Section 3. We provide details and licenses of the code at `https://github.com/socialfoundations/causal-features`. The original owners of the data are listed in Section 2.2 and Appendix E. The licenses, if available, are provided in Table 7.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide new code along with sufficient documentation at `https://github.com/socialfoundations/causal-features`.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.