Causal-structure Driven Augmentations for Text OOD Generalization

Anonymous Authors[1]

Abstract

In this work, we propose counterfactual data augmentation methods, guided by knowledge of the causal structure of the data, to simulate interventions on spurious features. Our main motivation is classifying medical notes, and we use these methods to learn more robust text classifiers. In prediction problems where the label is spuriously correlated with an attribute, and under certain assumptions, we show that this strategy is appropriate and can enjoy improved sample complexity compared to importance re-weighting. Pragmatically, we match examples using auxiliary data, based on diff-in-diff methodology, and use a large language model (LLM) to represent a conditional probability of text. Experiments on learning caregiver-invariant predictors of clinical diagnoses from medical narratives and on semi-synthetic data, demonstrate that our method improves out-of-distribution (OOD) accuracy.

1. Introduction

The reliance of Machine Learning models on spurious correlations can compromise safety and degrade performance in applications such as medical imaging (Zech et al., 2018; DeGrave et al., 2021), text classification (McCoy et al., 2019), and risk prediction systems (Caruana et al., 2015). Failures occur under distribution shift (Quinonero-Candela et al., 2008; Subbaswamy et al., 2019; Finlayson et al., 2021), which may result from differences in data recording protocols, shifts in the underlying population being monitored, or the way the model is being used. In this paper, we focus on text classification and explore how domain-informed use of language models can help us avoid such failures.

Consider a scenario where we want to make robust predictions about patients' conditions, probability of readmission, etc., using clinical narratives written in hospitals (Spyns, 1996; Zhou and Hripcsak, 2007). A common issue arises when patients with certain conditions are directed to specific caregivers in the hospital. When we train a predictor on data that exhibits a correlation between caregiver-specific style and clinical outcomes, the predictor may unintentionally rely on the style to make predictions. This leads to poor OOD generalization on data from unseen hospitals, due to changes in clinical practice (Finlayson et al., 2021).

In this work we develop *causally-driven data augmentation methods*, that leverage auxiliary data (e.g., time, document type, demographics) and domain knowledge (e.g. some traits, like demographics, may affect the caregiver a patient sees) to improve model robustness. Drawing on methods for learning invariant and shift-stable models (Peters et al., 2016; Magliacane et al., 2018; Arjovsky et al., 2019; Subbaswamy et al., 2019), and on the success of data augmentation in improving OOD generalization (Robey et al., 2021; Yao et al., 2022; Gao et al., 2023; Kaushik et al., 2019), our work lies at the intersection of these subfields (see short review of related work in Appendix A).

Intuitively, generating versions of clinical narratives as if they had been written by different caregivers (i.e. approximating counterfactual texts), de-correlates the writing style from the patient condition we wish to predict. However, it is difficult to achieve such data generation in practice and problem-specific traits must be taken into account (Kocaoglu et al., 2018). We draw on common causal inference methods to improve counterfactual estimation. While our approach can be applied to many modalities of data, in this work we focus on text classification and harness recent advances in LLMs. We present a formal setting motivating counterfactual augmentation for OOD generalization (§2), and our methods for counterfactual estimation (§3). Finally, we present our main experimental results (§4).

2. Problem Setting

Consider a classification problem with L classes where the label Y is spuriously correlated with a known attribute C (i.e. the correlation may change arbitrarily at test time, denoted by a red edge $C \leftrightarrow Y$ in Figure 1). This setting has been used previously to study learning with "shortcuts" (Makar et al., 2022) and spurious correlations (Veitch et al., 2021).

In our medical notes example, C is the caregiver writing the note and Y is the underlying condition we wish to diagnose.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

We denote the number of caregivers in our training data by K. For a hypothesis $h : \mathcal{X} \to \mathbb{R}^L$ and distribution P, the expected accuracy is denoted by $\mathcal{R}_P^{\ell_{01}}(h)$ and expected loss under a function $\ell : \mathbb{R}^L \times [L] \to \mathbb{R}$ by $\mathcal{R}_P^{\ell}(h)$. The data-generating process is depicted by the causal model in Figure 1, for our motivating example of clinical notes classification X is a vector representation of the clinical note and X^* is an unobserved sufficient statistic, representing all the relevant information about Y in the note that is unaffected by the writing style of the caregiver. Let us formally define this setting.

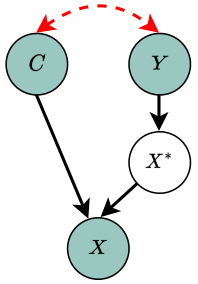

Figure 1: Prediction problem with a spuriously correlated attribute.

Definition 2.1. The set of distributions induced by interventions on a causal model with structure in Figure 1 is

$$\mathcal{P} = \Big\{ P(X \mid X^*, C) P(X^* \mid Y) P(Y) \tilde{P}(C \mid Y) \; : \; \tilde{P}(C \mid Y = y) \in \Delta^{K-1} \;\; \forall y \in [L] \Big\},$$

where all distributions other than $\tilde{P}(C \mid Y)$ are fixed. In a prediction problem with a spuriously correlated attribute, the learner is provided with a set $\{(\mathbf{x}_i, y_i, c_i)\}_{i=1}^N$ sampled i.i.d from $P_{\text{train}} \in \mathcal{P}$. We assume that $X^* = e(X)$ almost surely for some $e : \mathbb{R}^d \to \mathbb{R}^{d^*}$.

In this problem, once X^* is recovered no additional information from X is needed to predict Y. In clinical note classification, X^* represents all the information in the note about the patient conditions, unsullied by the writing style of caretaker C. To obtain $h^*(\mathbf{x})$ we will rely on risk minimization w.r.t a distribution where Y and C are uncorrelated. Consider the unconfounded distribution $P_\perp \in \mathcal{P}$ given by intervening on C, setting it independent of Y and uniformly distributed, $\tilde{P}(C \mid Y) = P_{\text{unif}}(C)$. An optimal classifier under $P_\perp$ is min-max optimal in the following sense.

Lemma 2.2. *For the prediction problem in Definition 2.1, the Bayes optimal classifier under the unconfounded distribution $P_\perp \in \mathcal{P}$ where C is uniformly distributed and independent of Y is $h^*(\mathbf{x}) = \arg\max_{y \in [K]} P_\perp(Y = y \mid X^* = e(\mathbf{x}))$. It is a minimizer of $\min_{h:\mathcal{X}\to[L]} \max_{P\in\mathcal{P}} \mathcal{R}_P^{\ell_{01}}(h)$ and $\mathcal{R}_P^{\ell_{01}}(h^*) = \mathcal{R}_{P_\perp}^{\ell_{01}}(h^*)$ for all $P \in \mathcal{P}$.*

Hence we would like to minimize risk w.r.t $P_\perp$ and we cannot do that directly via ERM since our training data is sampled from $P_{\text{train}} \neq P_\perp$. Instead we consider risk minimization over a dataset augmented with counterfactual instantiations of training data under different values of C.

Minimizing $\mathcal{R}_{P_\perp}$ via Counterfactual Data Augmentation. Returning to our motivating example, assume that we could obtain the clinical notes that would have been written if each patient had been seen by all possible caregivers $c \in [K]$, each writing their own version of the note $\mathbf{x}_i(c)$. Given these counterfactual clinical notes, we seek a hypothesis that minimizes the average loss over all such possible scenarios.

Definition 2.3. Consider a prediction problem with a spuriously-correlated attribute. For an example $\mathbf{x}_i$, we denote the counterfactual with attribute value $c \in [K]$ as derived from the corresponding causal model, by $\mathbf{x}_i(c)$. For estimates of the counterfactuals $\{\hat{\mathbf{x}}_i(c)\}_{i\in[N],c\in[K]}$ and hypothesis $h \in \mathcal{H}$, the counterfactually augmented empirical risk is $\widehat{\mathcal{R}}_{\text{aug}}^{\ell}(h) = (NK)^{-1} \sum_{i\in[N],c\in[K]} \ell\left(h\left(\hat{\mathbf{x}}_i(c)\right), y_i\right)$.

We use approximate counterfactuals $\hat{\mathbf{x}}_i(c)$ in our definition to highlight that in practice we cannot obtain a precise estimate of $\mathbf{x}_i(c)$. It is easy to show that in the ideal case where $\hat{\mathbf{x}}_i(x) = \mathbf{x}_i(c)$, the expected loss $\mathcal{R}_{\text{aug}}^{\ell}(h)$ where $N \to \infty$, satisfies $\mathcal{R}_{\text{aug}}^{\ell}(h) = \mathcal{R}_{P_\perp}^{\ell}(h)$ and the technique minimizes risk under $P_\perp$. Our main challenge is then to derive effective approximations for counterfactuals such as clinical notes under alternative writing styles.

3. Assumptions and Algorithms for Estimating Counterfactuals

Perfectly capturing writing style is a strong assumption. Even if we could perfectly model writing styles, we only observe a limited set of variables - the actual notes x, outcomes y, and assigned caregivers c. Other factors could influence what each caregiver would write. To alleviate this, we use auxiliary data M that is available during training, but might not be available in deployment.

As an example, consider two caregivers c and $\tilde{c}$, where a note $\mathbf{x}_i$ was written by $c_i = \tilde{c}$. We want to estimate what $\mathbf{x}_i(c)$, the note caregiver c would have written, might look like. To this end we learn a model $\tau_c(\cdot)$ that takes data and generates a note in caregiver c's style. Now suppose caregiver c usually sees patients with high blood pressure and always includes blood pressure values in notes, while $\tilde{c}$ rarely does. A naive model $\hat{\mathbf{x}}_i(c) = \tau_c(\mathbf{x}_i)$ learned only from c's notes may fill in false blood pressure information, conflating that with c's style. Including vitals data like blood pressure, typically recorded in a patient's health record, provides additional context for our model. This extra information assists the model in achieving more accurate estimates.

Using auxiliary data for counterfactual augmentation. To make effective use of this data, we suggest that the input to the model $\tau_c : \mathcal{X} \times \mathcal{M} \to \mathcal{X}$ will include a baseline text to be edited and auxiliary data $\mathbf{m}$. Our main use of $\mathbf{m}$ is to match units that are similar in their auxiliary data. In our example these are things such as vitals and drug prescriptions, and also includes the label y since we usually would like to preserve it. We specify the construction of τ_c in the following subsection.

Algorithm 1 *CATO*

Input: Training set $\{(\mathbf{x}_i, y_i, c_i, \mathbf{m}_i)\}_{i=1}^N$, Hypothesis class $\mathcal{H}$, Version $\in \{(A), (B)\}$, **Optional** pre-treatment data $\{(\mathbf{x}_{\text{pre},i})\}_{i=1}^N$.
if Version = (A) **then**
  Get $\tau_c(\mathbf{m}, \mathbf{x})$ with preprocess (A)
  Get $\hat{\mathbf{x}}_i(c) = \tau_c(\mathbf{x}_{i,\text{pre}}, \mathbf{m}_i)\ \forall i \in [N]$
else
  Get $\tau_c(\mathbf{m}, \mathbf{x})$ with preprocess (B)
  Get $\hat{\mathbf{x}}_i(c) = \tau_c(\mathbf{x}_i, \mathbf{m}_i)\ \forall i \in [N]$
end if
Return: $h_{\text{aug}} \in \mathcal{H}$ that minimizes $\widehat{\mathcal{R}}^{\ell}_{\text{aug}}$.

Pre-process *CATO* (A)

Assume: $\mathbf{m}$ includes the label y and pre-treatment attribute c_{pre}. We are given $\{\mathbf{x}_{j,\text{pre}}\}_{j=1}^N$.
Set $\boldsymbol{\rho}(c_j, \mathbf{m}_j) = \mathbf{x}_j - \mathbf{x}_{j,\text{pre}}$ for $j \in [N]$.
Return $\tau_c(\mathbf{x}, \mathbf{m}) := \mathbf{x}_{\text{pre}} + \rho(c, \mathbf{m})$

Pre-process *CATO* (B)

Assume: $\mathbf{m}$ includes the label y.
Return: prompt $\tau_c(\mathbf{x}, \mathbf{m})$ that rewrites $\mathbf{x}$ in the style of matching examples, i.e. $\{\mathbf{x}_j : (\mathbf{m}_j, c_j) = (\mathbf{m}, c)\}$.

3.1. Implemented Methods

Our framework for estimating $\mathbf{x}_i(c)$, *CATO* (**C**ausal-structure Driven **A**ugmentations for **T**ext **O**OD Generalization), involves the use of an LLM to model the conditional probability distribution of text. Counterfactuals are formed by matching similar auxiliary data examples or manipulating texts' vector representations, as described below.

Prompting with matched examples. Our first estimation method in Algorithm 1(B) draws insights from matching (Rosenbaum and Rubin, 1983). We construct a prompt for an LLM, that given an original text $\mathbf{x}$ and a set of context notes, asks the LLM to rewrite $\mathbf{x}$ in their style. Now given text $\mathbf{x}$ with auxiliary data $\mathbf{m}$ that we wish to estimate with counterfactual value c (i.e. writing style), $\tau_c(\mathbf{x}, \mathbf{m})$ runs this prompt with context notes whose auxiliary data is similar to $\mathbf{m}$ and their attribute value equals the desired c.

Diff-in-diff estimation. The procedure we use for medical note generation relies on additional structure involving panel data (i.e. data collected over time intervals across several individuals). A clinical narrative is usually consisted of several notes taken over the course of a patient's visit, each may be written by a different caregiver. Prediction is made using the release note from the hospital whose embedding consists our features $\mathbf{x}$. For simplicity let us consider a single note $\mathbf{x}_{\text{pre}}$ taken prior to $\mathbf{x}$. Difference-in-difference (Card and Krueger, 1993; Abadie, 2005; Angrist and Pischke, 2009) estimation of causal effect is based on the parallel-trends, or constant effect assumption that two units i, j with similar pre-treatment conditions would have seen the same effect had they been assigned the same treatment (in our case, the caregiver). Hence we assume our auxiliary data $\mathbf{m}$ includes c_{pre}, the caregiver assigned pre-treatment.

Assumption 3.1 (constant effect)**.** Let $\mathbf{x}_{i,\text{pre}}$ be the pre-treatment features for unit i, and assume $\mathbf{m}_i$ includes the pre-treatment attribute $c_{i,\text{pre}}$. There exists a function $\boldsymbol{\rho} : [K] \times \mathcal{M} \to \mathcal{X}$ such that $\mathbf{x}_i(c) = \mathbf{x}_{i,\text{pre}} + \boldsymbol{\rho}(c, \mathbf{m}_i)$.

Under this assumption, to calculate $\mathbf{x}_i(c)$ we can use any unit j for which $\mathbf{m}_i = \mathbf{m}_j$ and has $c_j = c$ to estimate $\boldsymbol{\rho}(c, \mathbf{m}_i) = \mathbf{x}_j - \mathbf{x}_{\text{pre},j}$. The resulting estimation procedure is given in Algorithm 1(B) and illustrated in Appendix C.1.3.

3.2. Sample Complexity Comparison

Reasoning about counterfactuals with problem-specific domain knowledge is a considerable challenge, and a simple alternative to that relies on less stringent assumptions involves re-weighting the loss function (see e.g. Shimodaira (2000); Makar et al. (2022)).

Reweighting baseline. Intuitively, re-weighting samples from the uncorrelated distribution $P(Y, C) = P(Y)P(C)$ by setting for each example i a weight $w_i = P_{\text{train}}(Y = y_i)P_{\text{train}}(C = c_i)/P_{\text{train}}(Y = y_i, C = c_i)$ and minimizing the weighted empirical risk $\hat{\mathcal{R}}^{\ell}_{\mathbf{w}}(h) = \frac{1}{m}\sum_{i\in[m]} w_i \ell\left(h(\mathbf{x}_i), y_i\right)$. It can be proved that at the limit of infinite data the method learns a min-max optimal hypothesis, as it also effectively minimizes $\mathcal{R}^{l}_{P_\perp}$ (see (Makar et al., 2022)). Hence it may seem like we do not stand to gain much from using augmentations. However, by combining results from Cortes et al. (2010) and a bound we prove in Lemma B.2 (see Appendix B), we can reason about the respective sample complexities of these methods. For reweighting the sample complexity scales as $(d_{2,\text{train}}(Y, C) \cdot N)^{-1/2}$, where $d_{2,\text{train}}$ is the exponent of the 2-Rényi divergence which measures dependence between Y and C in the training data. However for counterfactual data augmentation the scale is $N^{-1/2} + d_1\left(P_{\text{train}}(\tau_c(X, M)), P\left(X(c)\right)\right)$, where the total variation divergence $d_1(\cdot, \cdot)$ measures how well the augmentation τ_c estimates counterfactuals. We gather that when the spurious correlation is strong, yet data augmentation is accurate, our method may enjoy improved performance. Please see Appendix B for details.

Additional baselines. Counterfactuals are not the only type of causal knowledge that may be leveraged for learning more stable models. Many data dependent penalty terms have been proposed to impose conditional independence constraints drawn from the causal structure of the problem. Theory on these methods usually shows improved OOD performance under infinite data (Arjovsky et al., 2019; Wald et al., 2021; Puli et al., 2022; Veitch et al., 2021). Our baselines include a method based on the Maximum-Mean Discrepency (MMD) from Makar et al. (2022) who show improved sample complexity under a linear hypothesis class.

4. Experiments

We empirically study the following questions: (1) Can *CATO* enhance OOD performance of downstream classifiers? (2) Does it surpass the combination of reweighting and invariance penalties? (3) Is it more effective than alternative augmentation techniques, thus demonstrating the usefulness of the causal graph? (4) How sensitive is *CATO* to quality of counterfactuals?

See Appendix C for further details about the experiments.

Baselines. We compare *CATO* to several baselines:

- Observational - Baseline model trained on the original data. *PubMED BERT* (Gu et al., 2021) for *clinical narratives*, logistic regression for *restaurant reviews*.
- Reweighting - Baseline model with sample reweighting.
- MMD - Baseline model with an MMD penalty.
- Naive Augmentations - Baseline model on a dataset that also includes augmentations, generated by prompting an LLM to create more examples.
- Conditional Augmentations - Augmentations are generated by matching on auxiliary data and prompting an LLM to create one example in the the style of the other.

4.1. Clinical Narratives

Data. We consider three representative clinical NLP tasks, *clinical condition* prediction, *note segmentation* and *demographic traits* identification[1], for which we have both ID and OOD data. We utilize several electronic health records (EHR), training on MIMIC-III (Johnson et al., 2016). and i2b2 competitions as our held-out hospital datasets.

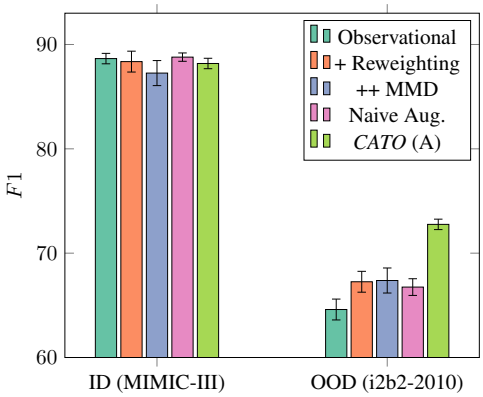

Figure 2: Results ($F1$ averaged across 5 runs) for predicting *clinical conditions* . *CATO* (A) outperforms on OOD data.

Clinical Condition Prediction. *Clinical condition* prediction is a concept extraction task focused on medical concepts in patient reports (Uzuner et al., 2011). Here we trained *PubMED BERT* models on a subset of MIMIC-III, labelled using the same annotation guidelines as in i2b2-2010, the OOD dataset the models are tested on. As can be seen in the Figure 2, in the ID setting only the naive augmentations improve performance slightly. In the OOD setting, all OOD methods help (*reweighting*, *MMD*, *CATO* (A)), but our causally-motivated augmentation approach is substantially better than the alternatives. On average (across 5 runs), *CATO* (A) improves precision above the baseline by more than 7% (absolute), and recall by more than 8%. The naive augmentation approach improves over the vanilla *PubMED BERT* model, but is outperformed by all OOD methods.

[1]See Appendix C for results on the *demographic traits* identification and *note segmentation*.

4.2. Restaurant Reviews

Data. We use the *CEBaB* dataset (Abraham et al., 2022), which consists of short restaurant reviews and ratings from OpenTable, including evaluations for food, service, noise, ambiance, and an overall rating. We construct two experimental settings: the original *CeBAB* dataset, and a modified version, denoted as *CeBAB*-Spurious, where there's a spurious correlation between training and deployment.

To construct *CeBAB*-Spurious, we leverage the availability of both the original and perceived ratings for each review in *CeBAB*. The original rating represents the reviewer's initial thoughts when writing the review, while the perceived rating indicates whether the review contains information about various restaurant attributes (e.g., food, service, noise, ambiance) and their associated sentiment. We utilize this unique data structure to capture reviewers' writing styles. Some reviewers are concise and provide limited descriptions, while others are more descriptive and include more information. To incorporate this variability, we introduce a new attribute called *food-mention* to signify the presence of food-related information in a review. If the perceived food rating is either negative or positive, we assign a value of 1 to the *food-mention* attribute; otherwise, it is set to 0. We subsample the data such that there is a correlation of 0.72 between *food-mention* and the outcome.

Method	*CeBAB*	*CeBAB*-Spur.
Observational	**0.85**	0.64
Reweighting	0.84	0.68
Naive Aug.	0.80	0.62
Conditional Aug.	0.84	0.70
CATO (B)	0.84	**0.75**

Table 1: Accuracy on *CeBAB* and *CeBAB*-Spurious. *CATO* (B) outperforms all baselines under a spurious correlation.

Results. As shown in Table 1, adding counterfactual augmentations leads to better OOD generalization, while naive data augmentation hurts model performance In line with the sample complexity argument in Section 3, conditional augmentation effectively doesn't add new data and therefore doesn't improve model performance.

References

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie 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*, pages 1721–1730, 2015.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Peter Spyns. Natural language processing in medicine: an overview. *Methods of information in medicine*, 35(04/05): 285–301, 1996.

Li Zhou and George Hripcsak. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40(2):183–202, 2007.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Neural Information Processing Systems (NeurIPS)*, pages 10869–10879, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Alexander Robey, George J. Pappas, and Hamed Hassani. Model-based domain generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Neural Information Processing Systems (NeurIPS)*, 2021. URL https://openreview.net/forum?id=JOxB9h40A-1.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022.

Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations. *arXiv preprint arXiv:2302.11861*, 2023.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Neural Information Processing Systems (NeurIPS)*, 34:16196–16208, 2021.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania, 1993.

Alberto Abadie. Semiparametric difference-in-differences estimators. *The review of economic studies*, 72(1):1–19, 2005.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Neural Information Processing Systems (NeurIPS)*, volume 23. Curran Associates, Inc., 2010.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Neural Information Processing Systems (NeurIPS)*, 34:2215–2227, 2021.

Aahlad Manas Puli, Lily H Zhang, Eric Karl Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=12RoR2o32T.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3 (1):1–23, 2021.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

Eldar D Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. *Neural Information Processing Systems (NeurIPS)*, 35:17582–17596, 2022.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain shifts. *arXiv preprint arXiv:2207.01603*, 2022.

Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In *Uncertainty in Artificial Intelligence*, pages 1546–1555. PMLR, 2021.

Mingzhang Yin, Yixin Wang, and David M Blei. Optimization-based causal estimation from heterogenous environments. *arXiv preprint arXiv:2109.11990*, 2021.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022a.

Amir Feder, Guy Horowitz, Yoav Wald, Roi Reichart, and Nir Rosenfeld. In the eye of the beholder: Robust prediction with causal user modeling. In *Neural Information Processing Systems (NeurIPS)*, 2022b.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

Yoav Wald, Gal Yona, Uri Shalit, and Yair Carmon. Malign overfitting: Interpolation can provably preclude invariance. *arXiv preprint arXiv:2211.15724*, 2022.

Adarsh Subbaswamy, Bryant Chen, and Suchi Saria. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, 10(1):64–89, 2022.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. Explaining the efficacy of counterfactually-augmented data. *arXiv preprint arXiv:2010.02114*, 2020.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

Rohan Jha, Charles Lovering, and Ellie Pavlick. Does data augmentation improve generalization in nlp? *arXiv preprint arXiv:2004.15012*, 2020.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1024. URL https://aclanthology.org/P17-1024.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. Textsettr: Label-free text style extraction and tunable targeted restyling. *arXiv preprint arXiv:2010.03802*, 2020.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*, 2021.

Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2021.

Maria Antoniak and David Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.148. URL https://aclanthology.org/2021.acl-long.148.

Xiaoling Zhou and Ou Wu. Implicit counterfactual data augmentation for deep neural networks. *arXiv preprint arXiv:2304.13431*, 2023.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. DoCoGen: Domain Counterfactual Generation for Low Resource Domain Adaptation. In *Proceedings of the 60th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2022.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29, 2017.

Özlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 2009.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6): 395–405, 2012.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015, 2016.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*, 2018.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*, 2019.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Dmitri Roussinov, Andrew Conkie, Andrew Patterson, and Christopher Sainsbury. Predicting clinical events based on raw text: from bag-of-words to attention-based transformers. *Frontiers in Digital Health*, 3:214, 2022.

Tom M Seinen, Egill A Fridgeirsson, Solomon Ioannou, Daniel Jeannetot, Luis H John, Jan A Kors, Aniek F Markus, Victor Pera, Alexandros Rekkas, Ross D Williams, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. *Journal of the American Medical Informatics Association*, 29(7): 1292–1302, 2022.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 2023.

Amir Feder, Itay Laish, Shashank Agarwal, Uri Lerner, Avel Atias, Cathy Cheung, Peter Clardy, Alon Peled-Cohen, Rachana Fellinger, Hengrui Liu, et al. Building a clinically-focused problem list from medical notes. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 60–68, 2022c.

Fan Zhang, Itay Laish, Ayelet Benjamini, and Amir Feder. Section classification in clinical notes with multi-task transformers. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 54–59, 2022.

Amir Feder, Danny Vainstein, Roni Rosenfeld, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. Active deep learning to detect demographic traits in free-form clinical notes. *Journal of Biomedical Informatics*, 107:103436, 2020.

Terence Tao. *An introduction to measure theory*, volume 126. American Mathematical Soc., 2011.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

Zihao Wang and Victor Veitch. A unified causal view of domain invariant representation learning. *arXiv preprint arXiv:2208.06987*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai,

and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Alexandra Pomares-Quimbaya, Markus Kreuzthaler, and Stefan Schulz. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC medical research methodology*, 19:1–20, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

OpenAI. Gpt-4 technical report, 2023.

Appendix

A. Related Work

Invariant and Shift-stable Learning. This paper contributes to the growing literature on invariant and shift-stable learning, which tackles the problem of learning models that generalizes across different distributions or settings. Invariant learning through feature pruning was pioneered by Peters et al. (2016), and has since been developed for variable selection (Magliacane et al., 2018; Heinze-Deml et al., 2018) and representation learning (Li et al., 2018; Arjovsky et al., 2019; Wald et al., 2021; Krueger et al., 2021; Puli et al., 2022; Makar et al., 2022; Jiang and Veitch, 2022). These methods have been applied in a range of domains, including natural science (Peters et al., 2016; Magliacane et al., 2018; Heinze-Deml et al., 2018), causal estimation (Shi et al., 2021; Yin et al., 2021), computer vision (Arjovsky et al., 2019; Krueger et al., 2021), and NLP (Veitch et al., 2021; Feder et al., 2022a;b). However, recent studies have highlighted limitations in many invariant learning approaches, particularly in achieving conditional independence (Kamath et al., 2021; Rosenfeld et al., 2020; Guo et al., 2021; Wald et al., 2022). Others have investigated learning of stable models by leveraging causal methods through techniques like graph-surgery (Subbaswamy et al., 2019; 2022), that come with generalization guarantees. Yet others have explored the advantages of data augmentation (Kaushik et al., 2019; 2020). In this work, we combine the latter two approaches to improve OOD generalization for text based classification.

Counterfactually Augmented Data. To learn invariant predictors, a popular and straightforward approach is *data augmentation*: construct counterfactual instances, and incorporate them into the training data. These counterfactuals involve perturbations to confounding factors (Garg et al., 2019), or to the label (Kaushik et al., 2019; 2020; Jha et al., 2020). Counterfactual examples can be generated through manual editing, heuristic keyword replacement, or automated text rewriting (Kaushik et al., 2019; Gardner et al., 2020; Shekhar et al., 2017; Garg et al., 2019; Feder et al., 2021; Zmigrod et al., 2019; Riley et al., 2020; Wu et al., 2021; Mao et al., 2021). Manual editing is accurate but expensive, while keyword-based methods can be limited in coverage and difficult to generalize across languages (Antoniak and Mimno, 2021). Generative approaches offer a balance of fluency and coverage (Zhou and Wu, 2023). Counterfactual examples help address causal inference's missing data issues, but generating meaningful counterfactuals is challenging (Calderon et al., 2022). Our work uses causal auxiliary data structure and LLMs to create plausible counterfactuals, enhancing OOD performance.

Clinical Notes. Clinical notes are the backbone of electronic health records, often containing vital information not observed in other structured data Kreimeyer et al. (2017). Clinical NLP involves identifying this information, and standardized datasets and competitions exist for this purpose (Uzuner, 2009; Savova et al., 2010; Jensen et al., 2012; Ford et al., 2016; Zhu et al., 2018). Best performing approaches have leveraged transformer architectures both for token-level classification tasks (Peng et al., 2019; Yadav and Bethard, 2019; Si et al., 2019; Lee et al., 2020), and for using complete clinical records (Roussinov et al., 2022; Seinen et al., 2022). Recently, large language models (LLMs), similar to those we use to generate counterfactual notes, were shown to have clear potential for improving clinical NLP systems (Singhal et al., 2022; Ayers et al., 2023). In our experiments, we follow recent papers in clinical NLP addressing challenges of degraded performance across different hospitals (Feder et al., 2022c; Zhang et al., 2022; Feder et al., 2020).

B. Proofs of Formal Claims

Notation. We will use random variables C, Y, M, X with images $[K], \mathcal{Y} = [L], \mathcal{M}, \mathcal{X}$ respectively in our probabilistic causal models. For a function $\tau_c : \mathcal{X} \times \mathcal{M} \to \mathcal{X}$, and measure P over sets in $\mathcal{X} \times \mathcal{M}$, we denote by $\tau_{c,*}P(X, M)$ the pushforward measure (Tao, 2011, §1.4). $\tau_c(\cdot)$ will be used to refer to the c-th coordinate of the output of a function $\tau : \mathcal{X} \times \mathcal{M} \to \mathcal{X}^K$. The notation $\mathcal{H}$ will be used for hypothesis classes where $h : \mathcal{X} \to \mathcal{Y}$ for any $h \in \mathcal{H}$. The $0-1$ loss $\ell_{01} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ is given by $\ell_{01}(\hat{y}, y) = 1_{\hat{y} \neq y}$. For a node V in a causal graph we will use $pa(V)$ for its causal parents.

For completeness we rewrite the definition of our data generating process from the main paper, this time adding the auxiliary data M into our model.

Definition 2.1. Consider a probabilistic causal model with endogenous random variables X, X^*, Y, C, M taking on values in $\mathcal{X}, \mathcal{X}^*, [L], [K], \mathcal{M}$ and exogenous independent random variables (Peters et al., 2017) $N_X, N_{X^*, N_Y, N_C, N_M}$, where the induced graph is a DAG that satisfies the following,

- Y is d-separated from X by X^*, C, M and also by X^*, C.
- Y, X^* are not descendants of C.

An anti-causal prediction problem with a spuriously-correlated attribute is a set of distributions $\mathcal{P}$ obtained by all interventions on C that replaces the distribution of exogenous noise N_C, mechanism $f_C(pa(C), N_C)$ with another mechanism (i.e. a measurable function $\tilde{f}(pa(C), N_C)$), or sets a fixed value (i.e. $do(C = c)$). Under the settings of this problem, a learner is provided with a set $\{(\mathbf{x}_i, y_i, c_i)\}_{i=1}^N$ sampled i.i.d from $P_{\text{train}} \in \mathcal{P}$.

We denote by $P_\perp \in \mathcal{P}$ the distribution obtained by intervening on C and setting it to a uniform distribution, i.e. $P_\perp(X, X^*, Y, C, M) = K^{-1} \sum_{c \in [K]} P(Y, X, X^*, M \mid do(C = c))$. Note that the problem described by Figure 1 and Definition 2.1 of the main paper is a special case of this setting where M is discarded, and $P_\perp$ coincides with setting $\tilde{P}(C \mid Y)$ to a uniform distribution.

Recall our assumption about perfect recovery of X^*.

Assumption B.1. For an anti-causal prediction problem with a spuriously correlated attribute, we assume that $X^* = e(X)$ a.e. for some $e : \mathcal{X} \to \mathcal{X}^*$.

Under these conditions $h(\mathbf{x}) = \arg\max_{y \in [L]} P_\perp(Y = y \mid X = \mathbf{x})$ is an optimal risk-invariant predictor as described below.

Lemma 2.2. For the prediction problem in Definition 2.1, the Bayes optimal classifier under the unconfounded distribution $P_\perp \in \mathcal{P}$ where C is uniformly distributed and independent of Y is $h^*(\mathbf{x}) = \arg\max_{y \in [K]} P_\perp(Y = y \mid X^* = e(\mathbf{x}))$. It is a minimizer of $\min_{h:\mathcal{X} \to [L]} \max_{P \in \mathcal{P}} \mathcal{R}_P^{\ell_{01}}(h)$ and $\mathcal{R}_P^{\ell_{01}}(h^*) = \mathcal{R}_{P_\perp}^{\ell_{01}}(h^*)$ for all $P \in \mathcal{P}$.

Proof. Assume $P_{\text{train}} \in \mathcal{P}$ is the distribution from which our training data is obtained. We will show that any hypothesis satisfying $h(X) = g \circ e(X)$ for some $g : \mathcal{X}^* \to \mathcal{Y}$ (i.e. that only depends on X^*) achieves the same risk over all $P \in \mathcal{P}$. To this end note that for such a hypothesis we have,

$$
\begin{aligned}
R_{P_{\text{train}}}^{\ell_{01}}(h) &= \int \ell_{01}(h(X), Y) P_{\text{train}}(X \mid Y, C, X^*, M) P_{\text{train}}(Y, C, X^*, M) dX^* dX dY dC dM \\
&= \int \ell_{01}(g \circ e(X), Y) P_{\text{train}}(X \mid C, X^*, M) P_{\text{train}}(Y, C, X^*, M) dX^* dX dY dC dM \\
&= \int \ell_{01}(g(X^*), Y) P_{\text{train}}(X \mid C, X^*, M) P_{\text{train}}(Y, C, X^*, M) dX^* dX dY dC dM \\
&= \int \ell_{01}(g(X^*), Y) P_{\text{train}}(X^*, Y) dX^* dY \\
&= \int \ell_{01}(g(X^*), Y) P(X^*, Y) dX^* dY.
\end{aligned}
$$

The first line writes down the expected risk explicitly, the second removes conditioning on Y in the distribution on X since we assumed Y is d-separated from X by C, X^*, M. In the third line we make it explicit that h depends on X^* alone, then we integrate out X, C, M. On the last line we remove the subscript train to denote that this distribution in fixed across $P \in \mathcal{P}$ as we assumed that X^*, Y are non-descendants of C (and members of $\mathcal{P}$ are obtained by interventions on C). Now for any $P \in \mathcal{P}$ we may repeat this derivation for $R_P^{l_{01}}(h)$ and we will obtain the same term (since $P(X^*, Y)$ are fixed regardless of the intervention applied in P, as we just argued), and we may conclude $R_{P_{\text{train}}}^{\ell_{01}}(h) = R_P^{\ell_{01}}(h)$.

Next to show that the Bayes optimal classifier over $P_\perp$ is the min-max optimal classifier w.r.t $\mathcal{P}$. Consider the interventional distribution where C is set to some fixed value $c \in [K]$, i.e. $P(X, X^*, Y \mid do(C = c))$. Under the graph we obtain from this intervention, Y is d-separated from X given X^*. Hence,

$$
\begin{aligned}
P(Y \mid X = \mathbf{x}, do(C = c)) &= \int_{X^*} P(Y \mid X^*, X = \mathbf{x}, do(C = c)) P(X^* \mid X = \mathbf{x}, do(C = c)) dX^* \\
&= P(Y \mid X^* = e(\mathbf{x}), X = \mathbf{x}, do(C = c)) \\
&= P(Y \mid X^* = e(\mathbf{x}), do(C = c)),
\end{aligned}
$$

where the first equality holds since $X^* = e(X)$ and the second from d-separation. Hence the Bayes optimal classifier under $P(Y, X \mid do(C = c))$ is $h^*(\mathbf{x}) = g \circ e(\mathbf{x}) = \arg\max_{y \in [L]} P(Y = y \mid e(\mathbf{x}), do(C = c))$. As argued earlier, since Y, X^* are non-descendants of C, it holds that $P(Y \mid e(X), do(C = c))$ is fixed across all $c \in [K]$. Hence $h^*(\mathbf{x})$ is the Bayes optimal classifier for all such interventional distributions and also for $P_\perp(X, Y) = \frac{1}{K} \sum_{c \in [K]} P(X, Y \mid do(C = c))$, and from our earlier discussion it is risk-invariant, i.e. $R_{P_\perp}^{\ell_{01}}(h^*) = R_P^{\ell_{01}}(h^*)$ for all $P \in \mathcal{P}$, which also means $\max_{p \in \mathcal{P}} R_P^{\ell_{01}}(h^*) = R_{P_\perp}^{\ell_{01}}(h^*)$. It is the min-max optimal classifier w.r.t $\mathcal{P}$ since any $h \neq h^*$ will have $\max_{p \in \mathcal{P}} R_P^{\ell_{01}}(h) \geq R_{P_\perp}^{\ell_{01}}(h) \geq R_{P_\perp}^{\ell_{01}}(h^*)$. □

Next we turn to prove a bound on sample complexity of counterfactual data augmentations. In the following lemma, $d_1(\tau_{c,*}(P_{\text{train}}(X,M)) \mid P(X(c)))$ is a distance between the true distribution over counterfactual instances $P(X(c))$ and our augmented data $\tau_{c,*}(P_{\text{train}}(X,M))$.[2] Divergences other than total-variation can be used, resulting in tighter bounds, see e.g. Ben-David et al. (2010).

Lemma B.2. *Consider an anti-causal prediction problem with a spuriously-correlated attribute (Definition 2.1), a measurable function $\tau : \mathcal{X} \times \mathcal{M} \to \mathcal{X}^K$, and let $d_1(P,Q)$ denote the total variation distance between two distributions P, Q. Further let $\lambda_{aug} = \left[R^{\ell_{01}}_{aug}(h^*) + R^{\ell_{01}}_{P_\perp}(h^*)\right]$, where h^* is the optimal hypothesis w.r.t $\mathcal{R}^{\ell_{01}}_{P_\perp}$. For any $h \in \mathcal{H}$ and $\delta \in (0.5, 1)$, with probability at least $1-\delta$ over the draw of the training set,*

$$\mathcal{R}^{\ell_{01}}_{P_\perp}(h) \le \widehat{\mathcal{R}}^{\ell_{01}}_{aug}(h) + \sqrt{\frac{\log(1/\delta)}{N}} + K^{-1} \cdot \sum_{c\in[K]} d_1\left(\tau_{c,*}\left(P_{train}(X,M)\right), P\left(X(c)\right)\right) + \lambda_{aug}. \tag{1}$$

Proof. Our first step is to show that for any hypothesis $h \in \mathcal{H}$, if our augmentation process is exact in the sense that $\tau_c(X,M) = X(c)$ a.e., then the expected risk (i.e. risk taken over an infinitely large sample) on the augmented data coincides with that over the unconfounded distribution $P_\perp(X,Y) = P_{\text{unif}}(C)P(X,Y \mid do(C))$.

$$\begin{aligned}
\mathcal{R}^{\ell_{01}}_{\text{aug}}(h) &= \mathbb{E}_{P_{\text{train}}(C,Y,M,X)}\left[K^{-1}\sum_{c\in[K]} \ell_{01}\left(h\left(\tau_c(X,M)\right),Y\right)\right] \\
&= K^{-1}\sum_{c\in[K]} \mathbb{E}_{P_{\text{train}}(C,Y,M,X)}\left[\ell_{01}\left(h\left(X(c)\right),Y\right)\right] \\
&= K^{-1}\sum_{c\in[K]} \mathbb{E}_{P_{\text{train}}(C,Y,X)}\left[\ell_{01}\left(h\left(X(c)\right),Y(c)\right)\right] \\
&= K^{-1}\sum_{c\in[K]} \mathbb{E}_{P(Y,X|do(C=c))}\left[\ell_{01}\left(h\left(X\right),Y\right)\right] \\
&= \mathcal{R}^{\ell_{01}}_{P_\perp}(h).
\end{aligned} \tag{2}$$

To bound $\mathcal{R}^{\ell_{01}}_{\text{aug}}(h) - \widehat{\mathcal{R}}^{\ell_{01}}_{\text{aug}}(h)$ we note that $\{\mathbf{x}_i, y_i, \mathbf{m}_i\}_{i=1}^N$ are $i.i.d$ samples from a joint distribution, where we may consider the loss on each example as $K^{-1}\sum_{c\in[K]} \ell_{01}(h(\tau_c(\mathbf{x}_i,\mathbf{m}_i),y_i))$, then by standard results using the Hoeffding inequality, e.g. Mohri et al. (2018, Corollary 2.11), we get that for $\delta \in (0.5,1)$,

$$\mathcal{R}^{\ell_{01}}_{\text{aug}}(h) \le \widehat{\mathcal{R}}^{\ell_{01}}_{\text{aug}}(h) + \sqrt{\frac{\log(1/\delta)}{N}}. \tag{3}$$

Finally, to obtain our result consider any $c \in [C]$. Denote

$$\begin{aligned}
\mathcal{R}^{\ell_{01}}_{\text{aug},c}(h) &:= \mathbb{E}_{P_{\text{train}}(Y,M,X)}[\ell_{01}(h(\tau_c(X,M))Y)], \\
\mathcal{R}^{\ell_{01}}_{P_\perp,c}(h) &:= \mathbb{E}_{P(Y,X|do(C=c))}[\ell_{01}(h(X),Y)],
\end{aligned}$$

and for h^* denote $\mathcal{R}^{\ell_{01}}_{\text{aug},c}(h,h^*) := \mathbb{E}_{P_{\text{train}}(M,X)}[\ell_{01}(h(\tau_c(X,M)),h^*(\tau_c(X,M)))]$ and respectively for $\mathcal{R}^{\ell_{01}}_{P_\perp,c}(h,h^*) := \mathbb{E}_{P_\perp(X)}[\ell_{01}\left(h(X(c))\right),h^*(X(c)))]$. The rest of our derivation is along the lines of Ben-David et al. (2010, Theorem 2). We use the distance

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tau_{c,*}P_{\text{train}}(X,M), P(X(c))) = 2\sup_{g\in\mathcal{H}\Delta\mathcal{H}} \left|P_{\text{train}}(g(\tau_c(X,M))=1) - P(g(X(c))=1)\right|,$$

where $\mathcal{H}\Delta\mathcal{H} = \{g(\mathbf{x}) = 1_{h(\mathbf{x})\neq h'(\mathbf{x})} \mid h,h' \in \mathcal{H}\}$ is a set of binary hypotheses, i.e. functions that mark disagreements between hypotheses in $\mathcal{H}$. It is easy to see that $d_{\mathcal{H}\Delta\mathcal{H}}$ lower bounds d_1 which takes the supremum w.r.t all measurable subsets for the two measures, since the sets of inputs where $h(\mathbf{x}) = 1$ are contained in those subsets. Also from (Ben-David et al., 2010, Lemma 3) we have that for any hypotheses $h, h' \in \mathcal{H}$ it holds that

$$\left|R^{l_{01}}_{\text{aug},c}(h,h') - R^{l_{01}}_{P_\perp,c}(h,h')\right| \le \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(\tau_{c,*}P_{\text{train}}(X,M), P(X(c))\right).$$

[2] The notation $\tau_{c,*}(\cdot)$ denotes the pushforward measure. We note that in our implementation τ_c is data dependent and we ignore this dependence to enable a simple analysis.

Then following the proof in Ben-David et al. (2010, Theorem 2), where the first and third inequalities will rely on the triangle inequality for classification errors (Crammer et al., 2008), we may get:

$$\begin{aligned}
\mathcal{R}^{\ell_{01}}_{P_\perp,c}(h) &\le \mathcal{R}^{\ell_{01}}_{P_\perp,c}(h^*) + \mathcal{R}^{\ell_{01}}_{P_\perp,c}(h,h^*) \\
&\le \mathcal{R}^{\ell_{01}}_{P_\perp,c}(h^*) + \mathcal{R}^{\ell_{01}}_{\text{aug},c}(h,h^*) + [\mathcal{R}^{\ell_{01}}_{P_\perp,c}(h,h^*) - \mathcal{R}^{\ell_{01}}_{\text{aug},c}(h,h^*)] \\
&\le \mathcal{R}^{\ell_{01}}_{P_\perp,c}(h^*) + \mathcal{R}^{\ell_{01}}_{\text{aug},c}(h,h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(\tau_{c,*}P_{\text{train}}(X,M), P(X(c))\right) \\
&\le \mathcal{R}^{\ell_{01}}_{\text{aug},c}(h) + \mathcal{R}^{\ell_{01}}_{P_\perp,c}(h^*) + \mathcal{R}^{\ell_{01}}_{\text{aug},c}(h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(\tau_{c,*}P_{\text{train}}(X,M), P(X(c))\right) \\
&= \mathcal{R}^{\ell_{01}}_{\text{aug},c}(h) + \mathcal{R}^{\ell_{01}}_{P_\perp,c}(h^*) + \mathcal{R}^{\ell_{01}}_{\text{aug},c}(h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(\tau_{c,*}P_{\text{train}}(X,M), P(X(c))\right)
\end{aligned}$$

Finally, we note that $\mathcal{R}^{\ell_{01}}_{P_\perp}(h) = K^{-1}\sum_{c\in[K]}\mathcal{R}^{\ell_{01}}_{P_\perp,c}(h)$ and similarly we have that $\mathcal{R}^{\ell_{01}}_{\text{aug}}(h) = K^{-1}\sum_{c\in[K]}\mathcal{R}^{\ell_{01}}_{\text{aug},c}(h)$, hence applying the above inequality for all $c \in [K]$ and averaging we get:

$$\begin{aligned}
\mathcal{R}^{\ell_{01}}_{P_\perp}(h) &\le \mathcal{R}^{\ell_{01}}_{\text{aug}}(h) + \frac{1}{2}K^{-1}\sum_{c\in[K]} d_{\mathcal{H}\Delta\mathcal{H}}\left(\tau_{c,*}P_{\text{train}}(X,M), P(X(c))\right) + \lambda_{\text{aug}} \\
&\le \mathcal{R}^{\ell_{01}}_{\text{aug}}(h) + K^{-1}\sum_{c\in[K]} d_1\left(\tau_{c,*}P_{\text{train}}(X,M), P(X(c))\right) + \lambda_{\text{aug}}.
\end{aligned}$$

Combining with Equation (3) we get the desired result. □

Sample Complexity of Importance Reweighting. Recall that re-weighting sets for each example i a weight $w_i = P_{\text{train}}(Y = y_i)P_{\text{train}}(C = c_i)/P_{\text{train}}(Y = y_i, C = c_i)$ and minimizes the weighted empirical risk:

$$\hat{\mathcal{R}}^{\ell}_{\mathbf{w}}(h) = \frac{1}{m}\sum_{i\in[m]} w_i \ell\left(h(\mathbf{x}_i), y_i\right).$$

It can be proved that at the limit of infinite data the method learns a min-max optimal hypothesis, as it also effectively minimizes $\mathcal{R}^{l}_{P_\perp}$ (see (Makar et al., 2022)). Hence augmentations may not seem advantageous for identifying the correct hypothesis. However, reweighting can require a larger sample to identify the correct hypothesis, particularly when Y and C are highly correlated.[3]

To make this statement precise, we can apply the bounds from Cortes et al. (2010) and compare them with an upper bound that we will derive for our method in Lemma B.2. To this end, let us consider the exponent of the Rényi divergence as a measure of dependence between Y and C in the training data. The divergence is given by $d_{\alpha,\text{train}}(Y,C) = \left[\sum_{y\in[L],c\in[K]} P^{\alpha}_{\text{train}}(Y = y, C = c)/P^{\alpha-1}_{\text{train}}(Y = y)P^{\alpha-1}_{\text{train}}(C = c)\right]^{\frac{1}{\alpha-1}}$, and we may derive the following bound for a hypothesis $h \in \mathcal{H}$ and any $\delta \in [0,1]$:

$$\widehat{\mathcal{R}}^{\ell}_{\mathbf{w}}(h) \le \mathcal{R}^{\ell}_{P_\perp}(h) + \sqrt{\frac{2d_{2,\text{train}}(Y,C)\cdot\log(1/\delta)}{N}} + \frac{d_{\infty,\text{train}}(Y,C)}{N}. \tag{4}$$

A complementary lower bound on $\widehat{\mathcal{R}}^{l}_{\mathbf{w}}(h)$ can also be derived based on results in Cortes et al. (2010). Comparing this to Equation (1), as we generate better counterfactuals the term $d_1\left(\tau_{c,*}\left(P_{\text{train}}(X,M)\right), P\left(X(c)\right)\right)$ decreases and also $\mathcal{R}^{\ell_{01}}_{\text{aug}}(h)$ becomes similar to $\mathcal{R}^{\ell_{01}}_{P_\perp}(h)$ (see Equation (2)), hence the bound scales with $N^{-\frac{1}{2}}$, resulting in a gain of factor $d_{2,\text{train}}(Y,C)$ over the upper bound on $\widehat{\mathcal{R}}^{\ell_{01}}_{\mathbf{w}}(h)$ in Equation (4). We also show this through simulations in Appendix C.3.

C. Experimental Details

We provide here further details about the experimental setup, the datasets we use, hyperparameters chosen for training the models, and data splits. We also include additional experiments that were omitted from the main paper for brevity, including experiments on *demographic traits* and *note segmentation* in clinical narratives, and experiments on synthetic data.

[3] We remark that other works discuss the potential benefits of data augmentation for identification in other problem settings, e.g. (Wang and Veitch, 2022, Thm. 9) and (Gao et al., 2023).

Input (x)	Label (y)	ID Data	OOD Data	Spurious Feature (c)	auxiliary data (m)
Clinical Narratives	Condition Prediction Note Segmentation Demographic Traits	MIMIC-III	i2b2-2010 partner data i2b2-2006	Caregiver ID	Medications, Lab Results, Vitals
Restaurant Reviews	Restaurant Rating	CEBaB	CeBAB-Spurious	Food-mention	Service, Noise, Ambiance, Food
Synthetic Data	$\{0, 1\}$	Gaussians		$\{0, \cdots, 7\}$	–

Table 2: Description of all our tasks and their corresponding experimental setup.

C.1. Clinical Narratives

C.1.1. DATA

We describe here the *MIMIC-III i2b2-2006* and *i2b2-2010* datasets.

MIMIC-III. The *MIMIC-III* (Medical Information Mart for Intensive Care III) dataset is a large, publicly available database containing detailed and anonymized health-related data associated with over 40,000 patients who stayed in critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. *MIMIC-III* is a rich resource for researchers in various fields, such as medicine, data science, artificial intelligence, and healthcare analytics. The dataset contains a diverse range of data types, including demographics, vital signs, laboratory test results, medications, and clinical notes. The dataset contains over 2 million clinical notes contributed by over $3,500$ distinct healthcare professionals, including doctors, nurses, and other clinicians, with an average of 571 notes per author.

The notes in the *MIMIC-III* dataset come in various types, reflecting the diverse aspects of patient care and documentation in the intensive care setting. Some of the most common note types include:

- Nursing/Progress notes: These are daily notes written by nurses or other care providers, documenting the patient's progress, condition, and care provided.
- Radiology reports: Reports written by radiologists after interpreting medical imaging studies (e.g., X-rays, MRIs, CT scans).
- ECG reports: Reports documenting the interpretation of electrocardiogram results.
- Discharge summaries: Comprehensive summaries written by physicians when a patient is discharged from the hospital, outlining the patient's hospital course, treatments, and follow-up instructions.
- Physician consult notes: Notes written by specialists when consulted by the primary care team to provide their expert opinion on specific medical issues.
- Pharmacy notes: Notes documenting medication-related information, including dosing, administration, and potential drug interactions.
- Social work notes: Notes related to the patient's psychosocial status, including social and family support, living arrangements, and other relevant factors.

i2b2-2006. The i2b2 (Informatics for Integrating Biology and the Bedside) initiative is a collaborative effort that aims to develop new methods and tools for biomedical research. It focuses on the development of a scalable computational infrastructure that can be used to accelerate the translation of basic research findings into clinical applications. As part of this effort, i2b2 has hosted several shared tasks and challenges related to natural language processing and machine learning in healthcare.

In 2006, the first i2b2 challenge, known as the *i2b2-2006* challenge, was conducted, focusing on the identification of obesity and its comorbidities in discharge summaries. The dataset provided for the challenge contained 694 de-identified discharge summaries, which were randomly selected from the Research Patient Data Registry (RPDR) at Partners HealthCare. The dataset was divided into a training set of 514 discharge summaries and a test set of 180 discharge summaries. It is important

to mention that the *i2b2-2006* dataset is relatively small compared to the *MIMIC-III* dataset and does not provide detailed information about the number of distinct authors or the average number of notes per author.

However, the discharge summaries typically include various sections such as patient demographics, admission and discharge dates, admission diagnoses, hospital course, procedures, medications, and follow-up plans. These summaries are generally written by physicians at the time of patient discharge, providing an overview of the patient's medical condition, treatment received, and overall hospital stay.

i2b2-2010. The *i2b2-2010* challenge, also known as the i2b2/VA challenge, was a shared task organized by the i2b2 (Informatics for Integrating Biology and the Bedside) initiative in collaboration with the US Department of Veterans Affairs (VA). The challenge aimed to encourage the development of natural language processing (NLP) and machine learning techniques for extracting medical concepts from clinical narratives. Specifically, the *i2b2-2010* challenge focused on the identification of medical problems, tests, and treatments from free-text clinical records.

The dataset provided for the *i2b2-2010* challenge contained 826 de-identified clinical records, which were sourced from three different institutions: Partners HealthCare, the University of Pittsburgh Medical Center (UPMC), and the VA. The dataset was divided into a training set of 349 records and a test set of 477 records.

Similar to the *i2b2-2006* challenge, the *i2b2-2010* dataset is relatively small compared to the *MIMIC-III* dataset and does not provide detailed information about the number of distinct authors or the average number of notes per author. The clinical records in the dataset are composed of diverse note types, such as discharge summaries, progress notes, radiology reports, and pathology reports, contributed by physicians, nurses, and other healthcare professionals.

While the dataset does not provide specific information about the number of distinct authors, the fact that the notes were contributed by different types of healthcare professionals across multiple institutions increases the dataset's diversity, making it more representative of real-world clinical settings.

C.1.2. PUBMED BERT

In our clinical narratives experiments, we use *PubMED BERT* (Gu et al., 2021), a variant of of the original BERT model (Devlin et al., 2018), as our vanilla model. That is, all of the baselines and *CATO* all use it either for embedding clinical text or for predicting *conditions*, *demographic traits* and *note segments*.

PubMED BERT is a BERT-based (Bidirectional Encoder Representations from Transformers) model that has been pre-trained specifically on biomedical and scientific text data (Gu et al., 2021). The model leverages the BERT architecture, which is a transformer-based deep learning model that has gained significant attention in natural language processing (NLP) for its state-of-the-art performance across a wide range of tasks.

PubMED BERT is pre-trained on a large corpus of approximately 14 million biomedical abstracts from the PubMed database, which is a comprehensive repository of biomedical literature. By pre-training the model on domain-specific data, *PubMED BERT* is expected to have a better understanding of biomedical concepts, terminology, and language patterns compared to general domain models like BERT-base and BERT-large (Devlin et al., 2018).

The main advantage of using *PubMED BERT* for biomedical text mining tasks is its domain-specific knowledge, which can lead to improved performance and more accurate results when fine-tuned on various downstream tasks, such as named entity recognition, relation extraction, document classification, and question answering. Since *PubMED BERT* is pre-trained on a large corpus of biomedical text, it is better suited to capturing the unique language patterns, complex terminology, and the relationships between entities in the biomedical domain.

Hyperparameters for Fine-Tuning PubMED BERT on *MIMIC-III*. In our study, we leveraged a pre-trained *PubMED BERT* model and fine-tuned it on the *MIMIC-III* dataset. During pre-training, the model employed masked language modeling and next sentence prediction objectives. The architecture consisted of 12 layers, 768 hidden units, and 12 attention heads. For task-specific optimization, we used the following hyperparameters: a $3e-5$ learning rate with a linear warmup during the initial 10% of training steps, a batch size of 32, a maximum sequence length of 512 tokens, and a dropout rate of 0.1. The AdamW optimizer was applied with a 0.01 weight decay and a 1.0 gradient clipping threshold. To prevent overfitting, early stopping was based on validation loss and used a 3-epoch patience. The fine-tuning process ran for up to 20 epochs, unless early stopping criteria were met sooner.

The fine-tuning process was executed on a high-performance computing cluster with multiple NVIDIA Tesla V100 GPUs, each equipped with 32 GB of memory, using the *PyTorch* deep learning framework (Paszke et al., 2019). The dataset was preprocessed and tokenized using the *HuggingFace Transformers* library (Wolf et al., 2019).

C.1.3. Generating notes from counterfactual caregivers.

To generate augmentations, we select caregivers with multiple patients and notes for more than one patient. For each caregiver-patient pair where both their last progress note and discharge summary were written by that caregiver[4], we match them to similar patients having the same initial caregiver but a different one for their discharge summary. In matching, we select patients with similar medications and lab results (denoted as patient's auxiliary data m in Table 2). We then generate counterfactual discharge summaries for matched patients using Algorithm 1(A) and train the model using original data and generated counterfactuals.

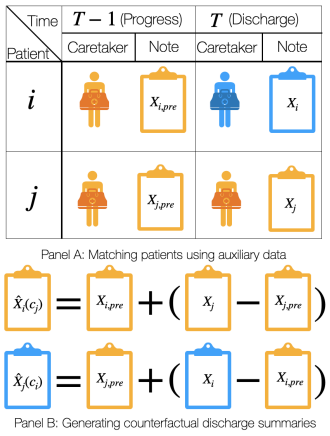

Figure 3: Generating counterfactual notes for patients with Algorithm 1(A).

C.1.4. *Demographic Traits* Detection

Demographic Traits detection is the task of identifying residual private information in the clinical note, after removing the known identifier types (names, ages, dates, addresses, ID's, etc.) (Feder et al., 2020). We train all models on a subset of *MIMIC-III* and test on *i2b2-2006*. Table 3 presents our results. While performance gains from the Causal Augmentation approach are not as large as in the other clinical NLP tasks, its is still the best method in terms of $F1$ score on out-of-distribution examples.

	ID (*MIMIC-III*)			OOD (*i2b2-2006*)		
	P	R	F1	P	R	F1
PubMED BERT	80.61	78.12	79.34	53.32	90.1	66.92
+ *Re-Weighting*	81.31	78.57	**79.92**	56.75	91.38	70.02
++ *MMD*	80.68	78.84	79.75	56.19	**91.49**	69.62
Naive Aug.	**81.45**	**79.35**	80.39	52.9	89.58	66.52
Causal Aug.	80.65	78.84	79.73	**59.76**	90.16	**71.88**

Table 3: Results (averaged across 5 runs) for predicting demographic traits from the text narratives on in-distribution and out-of-distribution data.

C.1.5. *Note Segmentation*

In this task, models need to recognize sections in free-form clinical notes (Pomares-Quimbaya et al., 2019). Given that section headers vary between hospitals, the models must discern sections based solely on the note content, excluding headers. As can be seen in Figure 4, similarly to *clinical condition* prediction, the diff-in-diff approach to augmentations (*CATO*

[4]During a patient's stay, progress notes capture its current state. When leaving the hospital, a discharge summary is written.

(A)) substantially improved OOD performance, and as expected does not help ID. The naive augmentations are the best performing method ID, but is again outperformed by all other methods OOD.

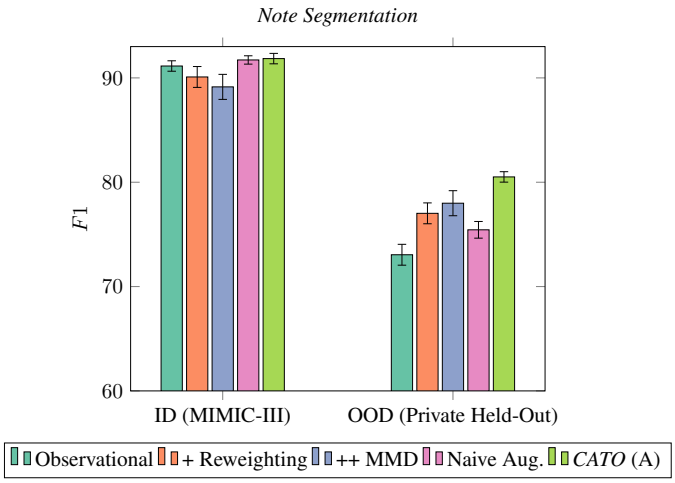

Figure 4: Results ($F1$ averaged across 5 runs) for clinical *note segmentation* from the text narratives. *CATO* (A) outperforms all baselines on OOD data.

C.2. Restaurant Reviews

Data. We use the *CEBaB* dataset (Abraham et al., 2022), which consists of short restaurant reviews and ratings from OpenTable, including evaluations for food, service, noise, ambiance, and an overall rating. For our experiments, we used the train-exclusive split of the dataset, which contains $1,755$ examples.

To analyze the data, we transformed the overall rating into a binary outcome. The original rating scale ranges from 1 to 5, and we classified a rating of 3 or higher as 1, and anything below as 0. We utilized a bag-of-words model with *CountVectorizer* and fitted logistic regression models from the *sklearn* library (Pedregosa et al., 2011).

To investigate these questions, we construct two experimental settings: the original *CeBAB* dataset, and a modified version, denoted as *CeBAB*-Spurious, where there's a spurious correlation between training and deployment.

The data is randomly split into a training set with $1,000$ examples and a test set with 755 examples. We explore two data augmentation schemes:

1. Naive data augmentation: This approach involves randomly selecting two reviews from the dataset and prompting *GPT-4* (OpenAI, 2023) to rewrite one restaurant review in the style of the other. By applying the naive augmentation, we obtain an additional $1,000$ training examples.

2. Conditional data augmentation : We match the ratings and sub-ratings in the reviews to create pairs. We then prompt *GPT-4* to rewrite one review to match the style of the other. Because not all pairs have matches in this case, the conditional data augmentation generates 926 augmentations. See Appendix C for details of the prompt.

Generating reviews with counterfactual food mentions. Following the counterfactual generation procedure in Algorithm 1, we generate counterfactual restaurant reviews conditional on food rating and overall rating. For each review, we first find a set of matched examples. We then select the subset that has different food-mention attribute and prompt *GPT-4* to rewrite. This results in $2,537$ augmentations. The counterfactual augmentation should capture what the reviews should look like had a reviewer been more/less concise. Following Algorithm 1, we generate counterfactual restaurant reviews conditional on food and overall ratings. We find matched examples for each review, select those with different food-mentions, and prompt a *GPT-4* to rewrite them, reflecting how the reviews would appear if the reviewer was more/less concise.

Prompt Example.

```
helper_prompt = """
you are a very helpful, diligent, and intelligent language model assistant,
your task to generate counterfactual restaurant reviews,
that is what the restaurant review would be if it is given a different rating.
You will be given an original restaurant review and a comparator review
Your task is to rewrite the original review, such that it will have the same
review score as the comparator review.
The rating is with respect to ambiance, food, noise, and service.
---- EXAMPLE INPUT - START -----

original_review: [],
original_ratings: [
rating_ambiance: score,
rating_food: score,
rating_noise: score,
rating_service: score
]

compare_reviews:[]
compare_ratings:[
rating_ambiance: score,
rating_food: score,
rating_noise: score,
rating_service: score
]


---- EXAMPLE INPUT - END -----
ANSWER FORMAT:
{
original_review: [],
original_score: [],
rewrite_review: [],
}

"""
```

C.3. Synthetic Data

To test sensitivity of *CATO* to quality of counterfactuals (Q#4), we generate synthetic data for a binary classification problem where $K = 8$ (cardinality of C). We sample $\tilde{P}(C \mid Y)$ to simulate varying degrees of the spurious correlation. Then we draw $\mathbf{x} = [\mathbf{x}^*, \mathbf{x}_{\text{spu}}]$ from a Gaussian distribution,

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_i^* \\ \mathbf{x}_{\text{spu},i} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{y_i} \\ \boldsymbol{\mu}_{c_i} \end{bmatrix}, \begin{bmatrix} \sigma^2 \mathbf{I}_{d^*} & 0 \\ 0 & \sigma^2_{\text{spu}} \mathbf{I}_{\mathbf{d_c}} \end{bmatrix} \right).$$

In this case $\hat{\mathbf{x}}_i(c)$ is obtained by adding $\mu_c - \mu_{c_i}$ to $\mathbf{x}_{\text{spu},i}$. To corrupt our augmentation, we instead add $\xi_i\,(\mu_c - \mu_{c_i})$ where ξ_i is drawn from a truncated Gaussian centered at $\lambda \in (0, 1)$. We train models with a fixed sample size (in the appendix we also examine varying sample sizes and additional types of corruption) and evaluate the trained models' accuracy on $P_\perp$ to examine the interplay between spurious correlation strength (measured by mutual information $I(Y; C)$), and counterfactual augmentation quality. As can be seen in Figure 5, corruptions degrade performance under stronger spurious correlations, though a strong corruption is required for reweighting to become preferable.

We study a binary classification problem where $K = 8$ (cardinality of C), and sample $\tilde{P}(C \mid Y)$ to simulate varying degrees

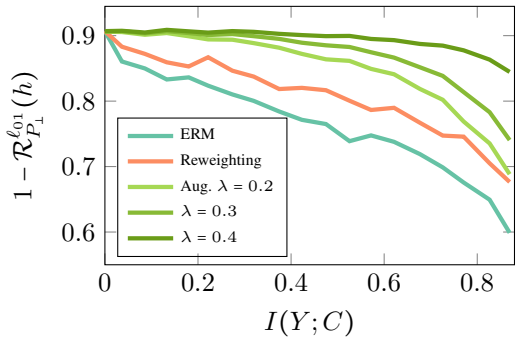

Figure 5: OOD accuracy ($1 - \mathcal{R}^{l_{01}}_{P_\perp}(h)$) and Y, C correlation strength ($I(Y;C)$). Even with substantial corruption ($\lambda = 0.2$) and strong correlation, augmentations outperform baselines.

of the spurious correlation (specifically, we draw). Then we draw $\mathbf{x} = [\mathbf{x}^*, \mathbf{x}_{\text{spu}}]$ from a Gaussian distribution,

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_i^* \\ \mathbf{x}_{\text{spu},i} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{y_i} \\ \boldsymbol{\mu}_{c_i} \end{bmatrix}, \begin{bmatrix} \sigma^2 \mathbf{I}_{d^*} & 0 \\ 0 & \sigma^2_{\text{spu}} \mathbf{I}_{\mathbf{d_c}} \end{bmatrix} \right).$$

In our simulations, we set $d^* = 10, d_{\text{spu}} = 300$ and $\sigma^2_{spu} = 0.05, \sigma = 0.01 d^*$ to make the max-margin classifiers depend on the spurious features. The parameters μ_{y_i}, μ_{c_i} are drawn uniformly from a sphere of norm $1/3$ and 60, respectively. For the corruptions of augmentations where we add $\xi_i(\mu_c - \mu_{c_i})$, the ξ_i variables are drawn from a truncated Gaussian centered at λ with standard deviation 0.1.

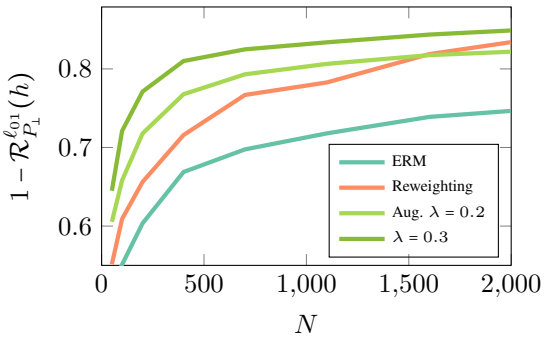

Figure 6: OOD accuracy ($1 - \mathcal{R}^{l_{01}}_{P_\perp}(h)$) for growing size of i.i.d training set N. We run 15 repetitions where $\tilde{P}(C \mid Y)$ are drawn randomly with correlation strength $I(Y;C) = 0.743 \pm 0.019$. With large amounts of data, the reweighting method approaches optimal performance and may outperform solutions based on corrupted data augmentation (e.g. it surpasses the more heavily corrupted data augmentation with $\lambda = 0.2$).

For the results in Figure 5 we set the number of training examples N at 600 and the distributions $\tilde{P}(C \mid Y)$ are sampled such that for each interval of size 0.05 between 0 and 0.9 for the values of $I(Y;C)$, we draw 30 instances within that interval. In Figure 6 we give results for another experiment where we plot curves for reweighting, ERM and corrupted augmentation under several values of N under a strong spurious correlation. We draw values for $\tilde{P}(C \mid Y)$ such that that $I(Y;C)$ is in $[0.7, 0.8]$ (mean 0.743 and standard deviation 0.019 with 15 repetitions). Considering the bounds in Equation (4) and the one in Lemma B.2, we expect that as N grows the reweighting method will approach optimal accuracy, while the augmentation methods will incur some bias due to the imperfect approximation of counterfactuals. In Figure 6 we see that the augmentation based methods begin to plateau, while the reweighting method gains larger improvement from the growing N, and outperforms the more heavily corrupted augmentation method ($\lambda = 0.2$) with a large enough training set. The best possible mean accuracy in this simulation is 0.934 (as we gather from training a classifier from $\mathbf{x}^*$ alone), the best accuracy obtained by the methods in the figure is 0.849 (Aug. $\lambda = 0.3$ with $N = 1999$).

D. Synthetic Note Examples

While it is not possible to share real notes from the *MIMIC-III* or any of the *i2b2* datasets, we can share how a typical note might look like by generating synthetic notes with LLMs. Here, we generate a synthetic note using *GPT-4,* and then generate a counterfactual note using our prompt:

Discharge Summary Note
Dr. Jane Doe
March 29, 2023
Patient Name: John Doe
Admission Date: March 25, 2023
Discharge Date: March 29, 2023
Diagnosis: Pneumonia
History of Present Illness: Mr. Doe is a 58-year-old male with a past medical history of hypertension who presented to the emergency department with a 3-day history of fever, cough, and shortness of breath. Chest X-ray showed bilateral infiltrates, and he was admitted for treatment of community-acquired pneumonia.
Hospital Course: The patient was started on intravenous antibiotics and supplemental oxygen. Over the course of his hospitalization, his oxygen requirements decreased, and his fever and cough resolved. He was able to tolerate a regular diet and was mobilized with physical therapy.
Consultations: Infectious disease was consulted, and the patient's antibiotics were adjusted based on culture results.
Discharge Medications:

1. Levofloxacin 750mg daily for 10 days
2. Acetaminophen 650mg every 6 hours as needed for fever or pain
3. Albuterol inhaler 2 puffs every 4 hours as needed for shortness of breath

Follow-up: The patient was instructed to follow up with his primary care physician in 1 week for a repeat chest X-ray and evaluation of his symptoms.
Discharge Instructions:

1. Finish the entire course of antibiotics as prescribed.
2. Use the albuterol inhaler as needed for shortness of breath.
3. Rest and avoid strenuous activity until cleared by your doctor.
4. Follow up with your primary care physician in 1 week.

Discharge Condition: The patient was discharged in stable condition with improved symptoms.

Here's a counterfactual discharge summary note written by a doctor with a different writing style:

Patient Name: John Doe
Admission Date: March 25, 2023
Discharge Date: March 29, 2023
Diagnosis: Pneumonia
HPI: John is a middle-aged man who came in with a fever and cough that he's had for three days. He was diagnosed with pneumonia and admitted.
Hospital Course: John was given antibiotics and oxygen. He slowly got better and was able to eat normally and move around more. He saw an infectious disease specialist who adjusted his treatment.
Consultations: Infectious disease saw John and changed his medicine.
Discharge Medications:

1. Levofloxacin 750mg once a day for 10 days
2. Acetaminophen 650mg every 6 hours as needed for fever or pain
3. Albuterol inhaler 2 puffs every 4 hours as needed for shortness of breath

Follow-up: Follow up with PCP in 1 week.
Discharge Instructions:

1. Finish your antibiotics.
2. Use the inhaler if you need it.
3. Rest and avoid heavy activity until you feel better.
4. Follow up with your doctor next week.

Discharge Condition: Stable, going home.

As can be seen from these examples, the counterfactual note is much more concise and to-the-point than the original example. The language used is more direct and less descriptive, and there is less detail provided about the patient's course of treatment.