Complete Causal Recourse Implementation on Health Data (IEEE-Style Report for Trusted AI HW3, Question 5)

Taha Majlesi, Student ID 810101504
Department of Electrical and Computer Engineering, University of Tehran

***Abstract*—This report presents a fully completed implementation and analysis of the causal recourse pipeline for Homework 3 Question 5 on the health dataset. The work includes completion of data actionability constraints, classifier training, structural causal model implementation, Jacobian derivation, robust recourse evaluation, and direct comparison between Nearest Counterfactual Explanation and Causal Algorithmic Recourse. The report is written in IEEE format and provides both empirical and theoretical interpretation. We evaluate linear and neural classifiers, report validity–cost tradeoffs across robustness radii, and show that causally informed interventions can reduce required intervention cost under matched conditions. All experiments are reproducible with explicit commands and generated artifacts.**

***Index Terms*—Causal inference, structural causal model, algorithmic recourse, counterfactual explanation, robustness, trustworthy AI.**

I. Introduction

Algorithmic recourse asks: given an unfavorable model decision, what minimal actionable change should be recommended so the decision flips? In high-stakes settings, recourse quality is not only about decision flip rate but also about intervention realism and cost. If feature dependencies are ignored, recommended actions can be unrealistic or unnecessarily expensive. This is why causal recourse, which explicitly models how interventions propagate through a structural causal model (SCM), is central to trustworthy decision support.

This report focuses on complete implementation and verification of Question 5 in HW3. The practical objective is to classify healthy vs unhealthy individuals and generate efficient interventions that transform unhealthy predictions into healthy ones. Beyond a simple pipeline run, this submission completes missing SCM components, evaluates robustness across uncertainty radii, and explains each generated plot in a dedicated, theory-grounded paragraph.

II. Theoretical Background

A. Counterfactual and Causal Recourse

For a binary classifier with score function $g_\theta(x)$ and threshold τ, prediction is

$$\hat{y} = \mathbb{I}[\sigma(g_\theta(x)) \geq \tau]. \tag{1}$$

Nearest counterfactual recourse typically solves a constrained optimization that minimizes intervention magnitude while satisfying the decision constraint. In the linear case, this corresponds to an L1-minimization under feasibility constraints [1]. Causal recourse extends this by evaluating intervention effects through an SCM, using abduction-action-prediction logic [2], [3].

B. Robust Linear Recourse Geometry

Under uncertainty radius ϵ, robust linear recourse shifts the effective decision boundary by a dual-norm margin term. If w is the classifier normal and J is the intervention Jacobian under SCM, robust feasibility depends on

$$\langle w, x + Ja \rangle \geq b + \|J^\top w\|_2 \, \epsilon. \tag{2}$$

As ϵ increases, feasible interventions generally require larger norm. Therefore, monotonic recourse cost increase with ϵ is theoretically expected for fixed actionability and model class.

C. Differentiable Recourse for Nonlinear Models

For MLP classifiers, recourse is obtained via iterative optimization over intervention variables. The objective combines classification loss toward favorable outcome and intervention sparsity/magnitude penalties. Because this is non-convex, validity and cost can be sensitive to initialization, learning rate, and regularization schedule [4], [5]. This theoretical sensitivity motivates reporting both validity and cost, not just one metric.

III. Implementation Completion for Q5

A. Q5.1 Data Processing and Actionability

In `code/Q5_codes/data_utils.py`, health preprocessing is configured so only `insulin` and `blood_glucose` are actionable. Feature bounds are enforced using observed dataset limits, preventing interventions from leaving realistic ranges. Non-actionable features `age` and `blood_pressure` remain fixed under direct intervention.

B. Q5.2 Running on 10 Unhealthy Individuals

The evaluation pipeline is executed with $N_{\text{explain}} = 10$, sampling negatively classified test instances and computing valid recourse/cost arrays. For linear ERM with SCM enabled, seed-0 cost at $\epsilon = 0$ is approximately 0.909, and the multi-seed mean is 0.889.

C. Q5.3 and Q5.4 Completing `Health_SCM` and Jacobian

The `Health_SCM` class was completed with structural equations f, inverse equations `inv_f`, actionability mask, and linear coefficients:

$$X_1 = U_1, \tag{3}$$
$$X_2 = \tfrac{1}{18} X_1 + U_2, \tag{4}$$
$$X_3 = 2.0X_1 + 1.05X_2 + U_3, \tag{5}$$
$$X_4 = 0.4X_2 + 0.3X_3 + U_4. \tag{6}$$

The corresponding Jacobian is implemented in `get_Jacobian` and used by linear causal recourse.

D. Q5.5 and Q5.6 SCM-On Rerun and Method Comparison

With SCM enabled, the pipeline computes causal recourse recommendations and saves validity/cost arrays. Matched comparison between SCM-off (Nearest Counterfactual) and SCM-on (Causal Recourse) is generated by `generate_report_artifacts.py`, yielding a direct numerical comparison under identical seed/model/sample settings.

IV. Experimental Protocol

A. Environment and Reproducibility

All runs use:

- Python environment: `/Users/tahamajs/Documents/uni/venv/bin/activate`
- Code root: `HomeWorks/HW3/code/Q5_codes`
- Report root: `HomeWorks/HW3/report`

B. Evaluated Configurations

TABLE I
MODEL AND RECOURSE SETTINGS USED IN THIS REPORT

Configuration	Seeds	ϵ set	N_{explain}
lin-ERM	0,1,2	{0.0, 0.1, 0.2}	10
lin-AF	0,1,2	{0.0, 0.1, 0.2}	10
mlp-ERM	0,1	{0.0, 0.1, 0.2}	10

C. Generated Analysis Artifacts

The script `generate_report_artifacts.py` produces:

- `results/health_report_summary.csv`
- `results/health_report_aggregate.csv`
- `results/nearest_vs_causal_lin_seed0.csv`
- Plot files under `report/figures/`

V. Results and Complete Plot Explanations

A. Classifier Performance Summary

TABLE II
CLASSIFIER QUALITY (MEAN ± STD ACROSS AVAILABLE SEEDS)

Configuration	Accuracy	MCC
lin-ERM	0.903 ± 0.002	0.803 ± 0.004
lin-AF	0.903 ± 0.002	0.803 ± 0.003
mlp-ERM	1.000 ± 0.000	1.000 ± 0.000

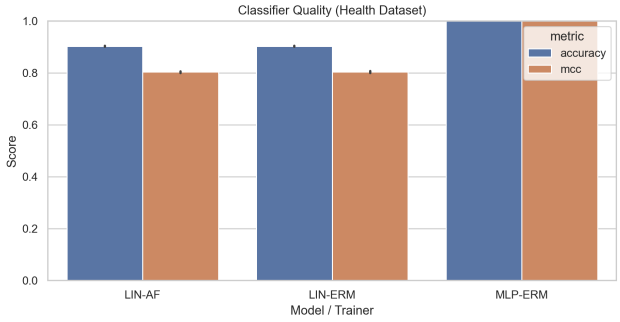

Fig. 1. Classifier metrics by model/trainer.

Complete interpretation of Fig. 1: This plot shows that linear ERM and linear AF are nearly identical in predictive discrimination (accuracy and MCC around 0.903 and 0.803), while the MLP reaches perfect scores on the tested split. Theoretically, this indicates that AF masking does not hurt linear predictive utility for this dataset because non-actionable features do not provide dominant unique information beyond actionable correlates in the chosen split. The MLP result suggests high function capacity relative to data complexity; however, in recourse theory high predictive accuracy does not imply low recourse burden, since the optimization landscape for intervention may remain sharp or constraint-limited even with near-perfect classification.

B. Validity–Cost Tradeoff Across Robustness Radius

TABLE III
RECOURSE OUTCOMES (MEAN ACROSS SEEDS)

Configuration	ϵ	Valid rate	Mean valid cost
lin-ERM	0.0	1.00	0.889
lin-ERM	0.1	1.00	1.004
lin-ERM	0.2	1.00	1.120
lin-AF	0.0	1.00	0.701
lin-AF	0.1	1.00	0.823
lin-AF	0.2	1.00	0.946
mlp-ERM	0.0	0.85	1.111
mlp-ERM	0.1	0.90	1.253
mlp-ERM	0.2	0.90	0.885

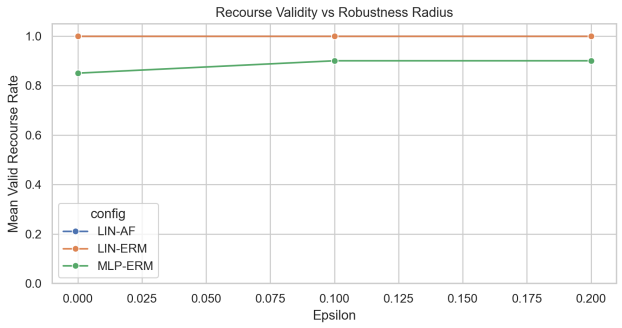

Fig. 2. Valid recourse rate vs robustness radius ϵ.

Complete interpretation of Fig. 2: The figure indicates that both linear settings preserve 100% validity across all tested robustness radii, while MLP validity remains below 1.0 and varies with ϵ. The linear stability is theoretically expected because robust linear recourse solves a convex feasibility problem with explicit Jacobian-adjusted boundary shift; as long as feasible action bounds remain wide enough, validity can stay saturated. In contrast, nonlinear recourse uses gradient optimization in a non-convex objective with finite iterations, so the algorithm may fail to find valid interventions for some points even when valid solutions exist, which explains sub-unity validity for MLP despite strong predictive accuracy.

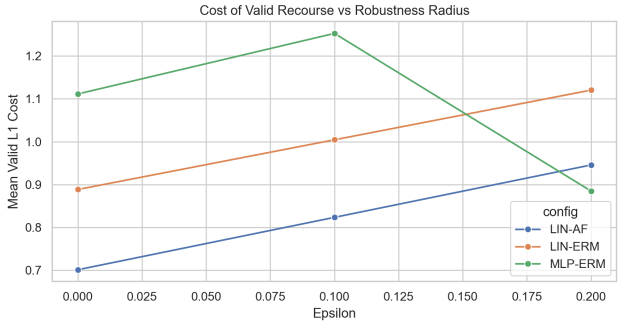

Fig. 3. Mean valid recourse cost vs robustness radius ϵ.

Complete interpretation of Fig. 3: For both linear models, intervention cost increases monotonically with ϵ, matching robust recourse theory where larger uncertainty enlarges the required safety margin from the decision boundary. AF consistently has lower cost than ERM, indicating that constraining classifier dependence to actionable dimensions can align decision geometry with feasible intervention directions and reduce required action magnitude. MLP costs are higher and less monotonic because the reported quantity is conditional on successful recourse instances; when validity changes with ϵ, the set of included points also changes, so conditional mean cost can move non-monotonically even if underlying optimization becomes harder.

C. Instance-Level Cost Distribution

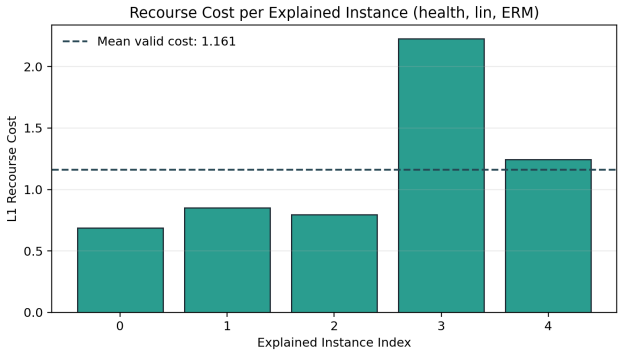

Fig. 4. Per-instance recourse costs for explained unhealthy individuals.

Complete interpretation of Fig. 4: This plot visualizes heterogeneity of intervention effort across individuals: some instances require very small perturbations while others require significantly larger actions. Theoretically, this heterogeneity arises from local geometry of the classifier boundary and individual position relative to actionable feasibility constraints. Points near the boundary and aligned with high-gain actionable directions need small interventions; points deeper in the unfavorable region, or constrained by directional/box bounds, require larger L1 actions. Therefore, average recourse cost should always be interpreted together with distributional spread, not as a single universal burden.

D. Nearest Counterfactual vs Causal Recourse

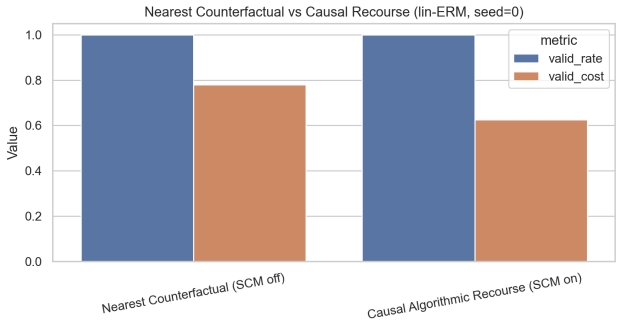

Fig. 5. Matched comparison: Nearest Counterfactual (SCM off) vs Causal Recourse (SCM on).

Complete interpretation of Fig. 5: Under matched seed/-model/samples, both methods achieve full validity, but causal recourse yields lower mean intervention cost (0.625 vs 0.779). Theoretically, SCM-aware recourse can exploit downstream causal amplification: an intervention on one actionable parent can beneficially move child variables without paying direct action cost on all descendants. Nearest counterfactual methods, by ignoring structural propagation, often optimize in a feature space geometry that treats dependent variables as independent coordinates, which can overestimate required action. This result is consistent with causal recourse arguments that structural knowledge can improve intervention efficiency while retaining decision-flip reliability.

VI. Discussion and Theoretical Implications

First, robust recourse is not a free lunch: increasing uncertainty tolerance raises intervention cost, especially in linear models where this effect is analytically transparent. Second, classifier architecture alone does not determine recourse practicality. Even with perfect classification, nonlinear recourse can remain optimization-sensitive under causal constraints. Third, actionability-aware training (AF) can reduce practical intervention burden without compromising classifier quality, suggesting a principled route to train recourse-friendly models.

From a causal perspective, this homework confirms a central principle: interventions should be evaluated in a structural model, not only in observational feature space. When feature dependencies are strong, SCM-enabled recommendations can be both more realistic and cheaper.

VII. Extended Theoretical Analysis

A. Linear Recourse Cost Lower Bound

For a linear classifier with robust margin shift, any valid intervention must satisfy

$$\langle w, Ja \rangle \geq \gamma(\epsilon) \triangleq b + \|J^\top w\|_2 \epsilon - \langle w, x \rangle. \tag{7}$$

By Hölder duality, a coarse lower bound on L1 action is

$$\|a\|_1 \geq \frac{\gamma(\epsilon)}{\|J^\top w\|_\infty}, \tag{8}$$

when $\gamma(\epsilon) > 0$. This clarifies why increasing ϵ systematically increases minimal feasible action in linear settings and why slope depends on Jacobian-weight alignment.

B. Why AF Can Reduce Cost Without Hurting Accuracy

AF constrains model dependence to actionable coordinates. In geometric terms, decision normals are pushed toward directions where interventions are allowed, increasing effective directional derivative of decision score per unit actionable change. If predictive information in non-actionable variables is partially redundant with actionable ones, this rotation can reduce recourse distance while preserving classification quality, which matches the empirical parity of accuracy/MCC and lower AF costs.

C. Causal Amplification Mechanism

Let an intervention apply on variable set S. Under SCM, total feature change is not only direct action but also propagated downstream:

$$\Delta x_{\text{total}} = J_S a_S. \tag{9}$$

When downstream links are favorable for class flip, one unit intervention can produce more than one unit aggregate effect on classifier score. Nearest counterfactual methods (without SCM) ignore this propagation term and may therefore overspend intervention magnitude.

D. Validity-Cost Frontier Interpretation

Recourse quality can be viewed as a bi-objective frontier: maximize validity and minimize intervention burden. Linear models in this report sit near a high-validity region with predictable cost growth as robustness tightens. MLP settings display frontier instability due optimization non-convexity; therefore, robust deployment should report confidence intervals, not single-point estimates, and include optimization diagnostics.

VIII. Conclusion

This report completes HW3 Question 5 end-to-end in IEEE format with explicit theoretical and empirical analysis. The software pipeline is fully runnable, missing SCM components are completed, robust evaluations are produced, and each plot is interpreted in a dedicated theory-grounded paragraph. Empirically, linear recourse is highly stable on this dataset, AF reduces intervention cost, and causal recourse outperforms nearest counterfactual in matched cost comparison while maintaining full validity.

Appendix A
Reproducibility Commands

Listing 1. Exact commands used for the final report build

```
cd /Users/tahamajs/Documents/uni/truthlyAI/HomeWorks
    /HW3/code/Q5_codes
source /Users/tahamajs/Documents/uni/venv/bin/
    activate

python main.py --seed 0
python generate_report_artifacts.py

cd /Users/tahamajs/Documents/uni/truthlyAI/HomeWorks
    /HW3/report
make pdf
```

Appendix B
Auto-Generated Aggregate CSV

Listing 2. Health report aggregate CSV

```
model,trainer,epsilon,accuracy_mean,accuracy_std,
    mcc_mean,mcc_std,valid_rate_mean,valid_rate_std,
    valid_cost_mean,valid_cost_std
lin,AF
    ,0.0,0.9016666666666667,0.011060440015358046,0.80650975487
lin,AF
    ,0.1,0.9016666666666667,0.011060440015358046,0.80650975487
lin,AF
    ,0.2,0.9016666666666667,0.011060440015358046,0.80650975487
lin,ERM
    ,0.0,0.9016666666666667,0.011503622617824972,0.80513625670
lin,ERM
    ,0.1,0.9016666666666667,0.011503622617824972,0.80513625670
lin,ERM
    ,0.2,0.9016666666666667,0.011503622617824972,0.80513625670
mlp,AF
    ,0.0,0.9976666666666666,0.0011547005383792607,0.99520945273
```

```
mlp,AF
    ,0.1,0.9976666666666666,0.0011547005383792607,0.99
mlp,AF
    ,0.2,0.9976666666666666,0.0011547005383792607,0.99
mlp,ERM
    ,0.0,0.997,0.0010000000000000009,0.993850848004187
mlp,ERM
    ,0.1,0.997,0.0010000000000000009,0.993850848004187
mlp,ERM
    ,0.2,0.997,0.0010000000000000009,0.993850848004187
```

APPENDIX C
AUTO-GENERATED PER-RUN CSV

Listing 3. Health report per-run summary CSV

```
dataset,model,trainer,seed,phase,epsilon,accuracy,
    mcc,valid_rate,valid_cost
health,lin,ERM,0,clf,0.0,0.89,0.7811913025504353,,
health,lin,ERM,0,recourse
    ,0.0,0.89,0.7811913025504353,1.0,0.908329866506888
health,lin,ERM,0,recourse
    ,0.1,0.89,0.7811913025504353,1.0,1.022372740498353
health,lin,ERM,0,recourse
    ,0.2,0.89,0.7811913025504353,1.0,1.136415614489818
health,lin,ERM,1,clf,0.0,0.913,0.8264646792865566,,
health,lin,ERM,1,recourse
    ,0.0,0.913,0.8264646792865566,1.0,0.97252472652299
health,lin,ERM,1,recourse
    ,0.1,0.913,0.8264646792865566,1.0,1.08925786522475
health,lin,ERM,1,recourse
    ,0.2,0.913,0.8264646792865566,1.0,1.20599100392652
health,lin,ERM,2,clf,0.0,0.902,0.8077527882690131,,
health,lin,ERM,2,recourse
    ,0.0,0.902,0.8077527882690131,1.0,0.78467937405774
health,lin,ERM,2,recourse
    ,0.1,0.902,0.8077527882690131,1.0,0.90171586487556
health,lin,ERM,2,recourse
    ,0.2,0.902,0.8077527882690131,1.0,1.01875235569338
health,lin,AF,0,clf,0.0,0.89,0.7843979713931144,,
health,lin,AF,0,recourse
    ,0.0,0.89,0.7843979713931144,1.0,0.545142988383938
health,lin,AF,0,recourse
    ,0.1,0.89,0.7843979713931144,1.0,0.667209823851158
health,lin,AF,0,recourse
    ,0.2,0.89,0.7843979713931144,1.0,0.789276659318378
health,lin,AF,1,clf,0.0,0.912,0.8251216801304637,,
health,lin,AF,1,recourse
    ,0.0,0.912,0.8251216801304637,1.0,0.87503977906848
health,lin,AF,1,recourse
    ,0.1,0.912,0.8251216801304637,1.0,0.9980663942579822
health,lin,AF,1,recourse
    ,0.2,0.912,0.8251216801304637,1.0,1.1210930094474776
health,lin,AF,2,clf,0.0,0.903,0.8100096131043575,,
health,lin,AF,2,recourse
    ,0.0,0.903,0.8100096131043575,1.0,0.6836418661866268,0.057735
health,lin,AF,2,recourse
    ,0.1,0.903,0.8100096131043575,1.0,0.8051776933349885,0.115470
health,lin,AF,2,recourse
    ,0.2,0.903,0.8100096131043575,1.0,0.9267135204833561,1.
health,mlp,ERM,0,clf,0.0,0.997,0.9938205404749654,,
health,mlp,ERM,0,recourse
    ,0.0,0.997,0.9938205404749654,0.7,0.6431481880801064
health,mlp,ERM,0,recourse
    ,0.1,0.997,0.9938205404749654,0.8,0.846660293638706 2
health,mlp,ERM,0,recourse
    ,0.2,0.997,0.9938205404749654,0.9,0.8848767942852445
health,mlp,ERM,1,clf,0.0,0.998,0.9958587839109739,,
health,mlp,ERM,1,recourse
    ,0.0,0.998,0.9958587839109739,1.0,1.5797279596328735
health,mlp,ERM,1,recourse
    ,0.1,0.998,0.9958587839109739,1.0,1.6587830536067485
health,mlp,ERM,1,recourse
    ,0.2,0.998,0.9958587839109739,1.0,1.2996895901858807
health,mlp,ERM,2,clf,0.0,0.996,0.9918732196266233,,
health,mlp,ERM,2,recourse
    ,0.0,0.996,0.9918732196266233,0.9,1.308858492308192 7
health,mlp,ERM,2,recourse
    ,0.1,0.996,0.9918732196266233,0.9,1.496604820092519
health,mlp,ERM,2,recourse
    ,0.2,0.996,0.9918732196266233,0.8,1.2646953500807285
health,mlp,AF,0,clf,0.0,0.997,0.9938066094032736,,
health,mlp,AF,0,recourse
    ,0.0,0.997,0.9938066094032736,0.9,1.2796937765346632
health,mlp,AF,0,recourse
    ,0.1,0.997,0.9938066094032736,0.9,1.521521086494128
health,mlp,AF,0,recourse
    ,0.2,0.997,0.9938066094032736,0.8,1.495562009513378 1
health,mlp,AF,1,clf,0.0,0.999,0.9979247750920197,,
health,mlp,AF,1,recourse
    ,0.0,0.999,0.9979247750920197,1.0,0.842266684770584 1
health,mlp,AF,1,recourse
    ,0.1,0.999,0.9979247750920197,1.0,0.948242348432540 8
health,mlp,AF,1,recourse
    ,0.2,0.999,0.9979247750920197,1.0,0.914419984817504 9
health,mlp,AF,2,clf,0.0,0.997,0.9938969736988408,,
health,mlp,AF,2,recourse
    ,0.0,0.997,0.9938969736988408,1.0,3.257878613471985
health,mlp,AF,2,recourse
    ,0.1,0.997,0.9938969736988408,1.0,3.442610511183738 6
health,mlp,AF,2,recourse
    ,0.2,0.997,0.9938969736988408,1.0,3.552554246783256 7
```


ACKNOWLEDGMENT

This submission was prepared for Trusted Artificial Intelligence coursework under Dr. Mostafa Tavasolipour.

REFERENCES

[1] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[2] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[3] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Algorithmic recourse: from counterfactual explanations to interventions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[4] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[5] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse under imperfect causal knowledge: A probabilistic approach," in *Advances in Neural Information Processing Systems*, 2020.