From Prediction to Structure: Causal Discovery for 30-Day Readmission Prediction

Amir Mohammad Karimi Mamaghan [1] **Yuyang Tao** [1] **M. Umar B. Niazi** [1 2] **Stefan Bauer** [3 4]
Karl Henrik Johansson [1 2]

Abstract

Hospital readmission within 30 days remains an important challenge in healthcare, with implications for care quality, discharge planning, and resource allocation. While machine learning models trained on electronic health records (EHRs) can accurately predict readmission risk, their explanations are typically associative and provide limited insight into the underlying problem structure. In this work, we combine predictive modeling and causal discovery to study 30-day readmission and mortality prediction on a preprocessed MIMIC-IV cohort. We first benchmark several tabular machine learning models and use SHAP-based attribution to identify the most influential predictors. We then apply three causal discovery methods to assess whether stable, clinically plausible dependency patterns emerge beyond purely predictive signals. Our results show that richer EHR feature sets substantially outperform a LACE-only baseline, with tree-based models achieving the strongest predictive performance. Across predictive, causal, and subgroup analyses, a consistent core set of variables centered on Physical Status, Length, LACE score, Comorbidity, and Age remains structurally important. These findings suggest that combining standard predictive models with causal discovery can provide a compact, interpretable structural view of readmission risk, offering an exploratory step toward identifying variables for future causal effect estimation and intervention-oriented analysis to reduce preventable readmissions.

[1]KTH Royal Institute of Technology, Stockholm [2]Digital Futures, Stockholm [3]Technical University of Munich [4]Helmholtz AI, Munich. Correspondence to: Amir Mohammad Karimi Mamaghan <amkm@kth.se>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Hospitals now generate large volumes of electronic health record (EHR) data during routine care, including demographics, diagnoses, medications, laboratory measurements, and physiological observations. These data provide an opportunity to study patient trajectories at scale and to support data-driven clinical decision-making. At the same time, EHR data are high-dimensional, heterogeneous, and often incomplete, which makes robust modeling and interpretation challenging (Ghassemi et al., 2020; Johnson et al., 2023). Nevertheless, they have become a central resource for machine learning research in healthcare, especially for risk prediction and retrospective outcome modeling (Ghassemi et al., 2020; Weissler et al., 2021).

Among the many prediction tasks studied on EHR data, 30-day hospital readmission has received particular attention because it is closely related to care quality, discharge planning, and healthcare utilization. In the United States, 30-day readmission measures are used in national quality monitoring and reimbursement programs, reflecting their operational and clinical importance (McIlvennan et al., 2015; James et al., 2023). Early identification of patients at high risk of readmission can support more targeted follow-up, better discharge coordination, and allocation of limited clinical resources (Leppin et al., 2014; Fischer et al., 2014). This makes readmission prediction a natural and practically important benchmark for studying EHR-based models.

However, most existing approaches to readmission prediction remain primarily correlational: they aim to predict which patients are likely to return, but provide limited insight into why the risk arises or which factors may be causally relevant. This distinction is important, since statistical association does not by itself imply causation, particularly in observational healthcare data where confounding, selection bias, and measurement artifacts are common (Shi & Norgeot, 2022; Sanchez et al., 2022; Karamitros et al., 2025). As a result, highly predictive models may still be difficult to translate into actionable clinical hypotheses. Causal discovery and related causal inference tools offer a complementary perspective by attempting to recover directed dependency structure from observational data, thereby moving

analysis beyond feature association alone (Shi & Norgeot, 2022; Kleinberg & Hripcsak, 2011).

In this paper, we combine predictive modeling and causal discovery to study 30-day readmission and mortality prediction on MIMIC-IV. We ask whether strong tabular predictors, SHAP-based variable selection, and multiple causal discovery methods can together recover a stable and clinically plausible structural view of readmission risk. Concretely, we first benchmark several tabular machine learning models on an independent and identically distributed (i.i.d.) EHR cohort and use SHAP-based attribution to identify the most influential predictors. We then apply three complementary causal discovery methods to fixed SHAP-selected feature subsets in order to examine whether a consistent dependency structure emerges beyond purely predictive signals. Our main contributions are as follows:

1. a benchmark of tabular machine learning models for 30-day readmission and mortality prediction on a preprocessed MIMIC-IV cohort;
2. a feature-level interpretation of predictive performance using SHAP-based attribution;
3. a comparative causal discovery analysis using NOTEARS, FCI, and DECI, on fixed SHAP-selected clinical variables, including robustness analysis across age groups.

Taken together, the paper is intended as a first step toward causally informed intervention design for reducing preventable hospital readmissions.

2. Related Work

Readmission prediction. Prior work on EHR-based readmission prediction has mainly focused on predictive modeling with structured variables, clinical notes, or temporal patient trajectories. Studies on MIMIC-III and MIMIC-IV have evaluated deep learning, recurrent, attention-based, and standard machine learning models for ICU readmission and related outcomes (Barbieri et al., 2020; Deng et al., 2022; Huang et al., 2019; de Sá et al., 2023). These works show that EHR data contain useful predictive signal, but their explanations are typically associational. Reproducible preprocessing pipelines such as COP-E-CAT also highlight the importance of cohort construction and feature extraction for downstream performance and interpretation (Mandyam et al., 2021).

Causal methods in healthcare. Causal inference and discovery are increasingly used in healthcare to move beyond correlation-based prediction toward explanation and intervention-oriented modeling (Kleinberg & Hripcsak, 2011; Shi & Norgeot, 2022; Sanchez et al., 2022). Prior work has applied causal graphical models and discovery algorithms to diagnostic decision support, biomarker analysis, neuroimaging, and psychiatric records (Richens et al., 2020; Shen et al., 2020; Miley et al., 2023). Closest to our setting, CARE-30 incorporates a learned DAG into a multimodal ICU readmission model, while Marafino et al. separate readmission risk prediction from treatment-effect estimation (Wang et al., 2023; Marafino et al., 2020). Unlike these works, we compare multiple causal discovery methods on the same EHR readmission cohort and examine whether they recover a stable structural view that complements standard predictive modeling.

3. Methodology

We study 30-day readmission and mortality prediction as a binary classification task on an i.i.d. EHR cohort, referring to it as 30-day readmission throughout. Our pipeline first benchmarks tabular predictive models, then uses SHAP-based attributions to select the most influential variables, and finally applies causal discovery to these fixed subsets to obtain compact graphs intended to characterize the structural neighborhood around readmission risk.

3.1. Dataset

We use MIMIC-IV v3.1 (Johnson et al., 2023), a large de-identified EHR database containing hospital- and ICU-level data from 2008 to 2019. We construct an i.i.d. cohort for 30-day readmission and mortality prediction by merging patient- and admission-level information, computing the LACE index (Van Walraven et al., 2010), and enriching the cohort with COP-E-CAT variables (Mandyam et al., 2021). We adapt COP-E-CAT preprocessing to a static setting by summarizing each patient stay as a single feature vector, followed by cleaning, categorical encoding, and retaining one record per patient. The final dataset contains 56,300 samples and 50 features; full cohort construction details and the complete variable list are provided in Appendix A.

3.2. Models

Predictive models. We benchmark six tabular prediction models for 30-day readmission: Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and TabPFNv2 (Hollmann et al., 2025). These models cover linear baselines, tree ensembles, gradient-boosted trees, and a recent tabular foundation model. We report standard classification metrics and use SHAP-based attributions (Lundberg & Lee, 2017) from the strongest-performing tree-based predictors to identify influential variables for causal analysis.

Causal discovery models. We compare three complementary causal discovery methods: nonlinear NOTEARS (Zheng et al., 2020), FCI with RCIT (Zhang, 2008), and DECI (Geffner et al., 2022). These represent score-based, constraint-based, and probabilistic approaches, respectively, and differ in their assumptions about latent confounding,

Table 1. Readmission prediction performance on LACE-only and full feature sets. We report F1-score (F1) and ROC AUC (AUC).

	LACE-only		All features	
Model	F1	AUC	F1	AUC
Random Forest	0.645	0.689	**0.836**	**0.907**
CatBoost	0.645	0.690	0.792	0.841
XGBoost	0.645	0.689	0.776	0.824
LightGBM	0.645	**0.690**	0.745	0.793
TabPFNv2	**0.656**	0.688	0.732	0.787
Logistic Regression	0.633	0.676	0.686	0.726

graph orientation, and structural uncertainty. We apply them to fixed SHAP-selected feature subsets rather than the full feature space, since causal discovery in higher-dimensional EHR settings becomes less stable, less interpretable, and substantially more computationally demanding. Full model and implementation details are provided in Appendix B.

4. Experiments

We evaluate the pipeline along three axes: predictive performance, feature-level interpretability, and inferred causal structure. We first benchmark tabular models for 30-day readmission prediction using both LACE-only features and the full EHR feature set (Section 4.1). We then use SHAP attributions from the strongest predictors to identify influential variables and define fixed top-5 and top-10 subsets for causal analysis (Section 4.2). Finally, we apply three causal discovery methods to the fixed top-10 subset and compare the resulting graphs (Section 4.3). Additional causal discovery and age-stratified robustness results are provided in Appendices C and D.

4.1. Readmission Prediction

We first benchmark the predictive performance of tabular ML models for 30-day readmission prediction under two feature configurations: (i) a clinical baseline using only the four LACE variables, and (ii) the full preprocessed EHR feature set with all 50 variables. We evaluate Logistic Regression, Random Forest, gradient-boosted tree models (XGBoost, LightGBM, CatBoost), and TabPFNv2, and report F1-score and ROC AUC on a held-out test set. We train the models on 80% of the data and evaluate them on the remaining 20%. For TabPFNv2, we use 10,000 randomly selected training samples due to the data-size limitation of the model version available at the time of experimentation.

Results. As shown in Table 1, the LACE-only setting yields modest and similar performance across models, with F1 around 0.63–0.66 and AUC around 0.68–0.69. In contrast, the full EHR feature set substantially improves performance for all models. Tree-based methods perform best: Random Forest achieves the strongest results with F1=0.836 and AUC=0.907, followed by CatBoost and XGBoost. These results show that richer EHR representations provide substantial predictive value beyond LACE alone, and that the predictive signal is robust across multiple non-linear models.

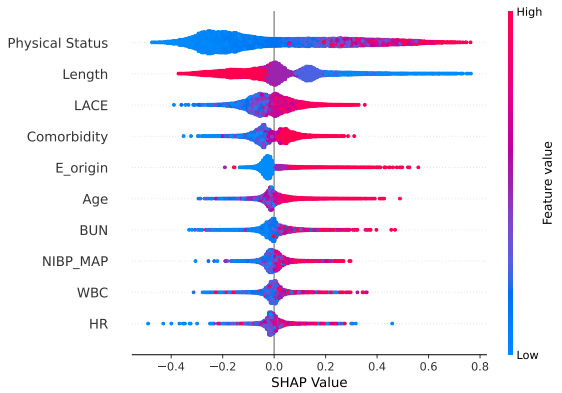

Figure 1. SHAP summary plot for the Random Forest model on the top 10 features. Features are ordered by mean absolute SHAP value.

4.2. Feature Importance Analysis

To better understand the behavior of the predictive models, we analyze SHAP values for the Random Forest model and present the resulting top-10 features in Figure 1. We report Random Forest for clarity because the top-performing tree-based models show highly similar feature rankings. The SHAP results reveal a stable set of clinically meaningful predictors that define the variable subsets used in the subsequent causal analysis.

Results. *Physical Status* is the most influential feature by a clear margin. Since this variable summarizes how often a patient exhibits abnormal clinical measurements during the hospital stay, its dominance suggests that overall clinical instability is a major driver of readmission risk. Higher values of *Physical Status* are associated with larger positive SHAP values, indicating that greater instability increases predicted readmission risk. Beyond this, *Length*, *LACE*, and *Comorbidity* also rank highly, showing that hospitalization duration, LACE score, and burden of chronic disease all contribute substantially to the model's decision process.

The remaining top features provide complementary physiological and utilization-related context. Higher *Age* is generally associated with increased readmission risk, while *BUN*, *WBC*, *NIBP_MAP*, and *HR* highlight the importance of renal function, inflammatory status, hemodynamic condition, and cardiovascular state. *E_origin* further indicates that prior emergency-related utilization carries a meaningful predictive signal. Overall, the SHAP analysis identifies a coherent top-10 set that spans general instability, disease burden, utilization history, and physiological state. These variables form the basis of the fixed top-10 and top-5 subsets used in

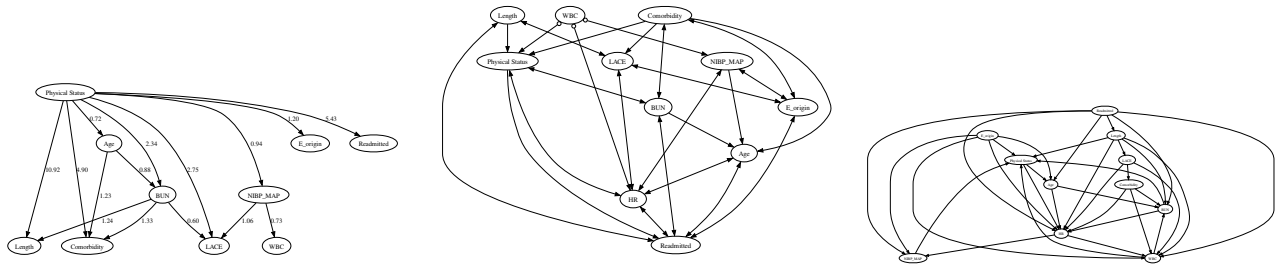

Figure 2. Causal discovery results on the top-10 SHAP-selected features for NOTEARS (left), FCI (middle), and DECI (right). Across methods, *Physical Status*, *Length*, *LACE*, *Comorbidity*, and *Age* remain central in the local neighborhood of readmission, while additional physiological variables such as *BUN*, *WBC*, *HR*, and *NIBP_MAP* provide complementary structure.

the causal experiments below.

4.3. Causal Discovery on Top Predictors

We next apply causal discovery to the fixed SHAP-selected top-10 and top-5 feature subsets and compare the resulting structures across NOTEARS, FCI, and DECI. We focus on the top-10 setting in the main text because it provides the best trade-off between interpretability and retained clinical detail: compared with the top-5 setting, it preserves additional physiological and utilization variables, while remaining substantially easier to interpret than a larger feature set. Since all three methods operate on the same selected variables, differences between them reflect differences in recovered dependency structure rather than differences in variable selection. The results of the top-5 setting are provided in Appendix C.

Results. The graphs are shown in Figure 2. Across all three methods, a consistent high-level picture emerges. First, *Physical Status* remains the most stable and prominent variable in the local neighborhood of *Readmitted*, consistent with the predictive SHAP analysis. Second, *Length*, *LACE*, *Comorbidity*, and *Age* repeatedly appear as central variables around readmission. Third, the top-10 setting preserves informative physiological variables such as *BUN*, *WBC*, *HR*, and *NIBP_MAP*, indicating that the recovered structure is not driven only by demographic or score-based features. Taken together, these results suggest that the strongest predictive variables are also structurally central in the inferred causal neighborhood around readmission.

The three methods differ mainly in density and orientation confidence rather than in which variables appear important. NOTEARS yields the clearest directed structure and is easiest to interpret qualitatively. FCI is more conservative in orientation, as expected with possible latent confounding, but preserves the same local neighborhood around readmission. DECI produces the densest graph, yet still highlights the same core variables. This cross-method agreement is one of the main findings of the paper: although edge-level details vary, the central readmission structure remains stable across substantially different causal discovery approaches. The corresponding top-5 pooled results and additional age-stratified NOTEARS analyses are provided in Appendices C and D.

5. Conclusion and Limitations

We study 30-day readmission prediction on a preprocessed MIMIC-IV cohort by combining tabular machine learning, SHAP-based feature attribution, and causal discovery. Richer EHR feature sets substantially outperform a LACE-only baseline, with tree-based models achieving the strongest predictive performance. SHAP analysis identifies a clinically meaningful set of influential variables centered on *Physical Status*, *Length*, *LACE*, *Comorbidity*, and *Age*, while multiple causal discovery methods recover a stable local structure around readmission despite differences in graph density and edge orientation. These results suggest that predictive modeling and causal discovery can provide a compact, hypothesis-generating structural view of readmission risk and help identify candidate variables for future causal effect estimation, counterfactual analysis, and intervention design.

This study has several important limitations. First, all analyses are based on a single observational EHR dataset, which limits external validity and prevents strong conclusions about generalization across hospitals or patient populations (Ghassemi et al., 2020; Finlayson et al., 2021). Second, causal discovery from observational healthcare data is fundamentally limited by potential unmeasured confounding, selection effects, and violations of assumptions such as acyclicity, causal sufficiency, or faithfulness (Spirtes et al., 2000; Pearl, 2009; Peters et al., 2017; Shi & Norgeot, 2022). Furthermore, compressing temporal EHR trajectories into static, i.i.d. feature vectors inherently discards time-dependent dynamics that could otherwise clarify true causal directions. For this reason, the recovered graphs should be interpreted as hypothesis-generating structures rather than definitive causal graphs.

Third, our graph assessment is primarily qualitative. In the absence of a validated ground-truth graph and interventional data, standard causal discovery metrics are not directly useful in this setting, and without additional assumptions, the underlying causal graph is non-identifiable (Spirtes et al., 2000; Pearl, 2009; Peters et al., 2017). We therefore rely on cross-method consistency, overlap with predictive feature importance, subgroup robustness, and clinical plausibility as our main evaluation criteria. Finally, the present study stops at structural analysis: it does not estimate intervention effects, evaluate counterfactuals, or learn decision policies. Addressing these limitations will be necessary before the framework can be translated into a clinically actionable intervention design.

Acknowledgements

This work is based on Yuyang Tao's master's thesis at KTH Royal Institute of Technology. The work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation, and by the Helmholtz Foundation Model Initiative, supported by the Helmholtz Association. Additional support was provided by the Swedish Research Council Distinguished Professor Grant 2017-01078 and the Knut and Alice Wallenberg Foundation Wallenberg Scholar Grant. The computations were enabled by the Berzelius resource, provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

Impact Statement

This paper presents an observational analysis of machine learning and causal discovery methods for readmission modeling from electronic health records. The main expected impact is methodological, as a step toward more interpretable and causally informed decision-support tools. In its current form, the work has no direct clinical impact; it provides a hypothesis-generating analysis that may inform future studies on causal effect estimation, intervention design, and clinical decision support.

References

Barbieri, S., Kemp, J., Perez-Concha, O., Kotwal, S., Gallagher, M., Ritchie, A., and Jorm, L. Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk. *Scientific reports*, 10 (1):1111, 2020.

de Sá, A. G., Gould, D., Fedyukova, A., Nicholas, M., Dockrell, L., Fletcher, C., Pilcher, D., Capurro, D., Ascher, D. B., El-Khawas, K., et al. Explainable machine learning for ICU readmission prediction. *arXiv preprint arXiv:2309.13781*, 2023.

Deng, Y., Liu, S., Wang, Z., Wang, Y., Jiang, Y., and Liu, B. Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients. *Frontiers in Medicine*, 9:933037, 2022.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Fischer, C., Lingsma, H. F., Marang-van De Mheen, P. J., Kringos, D. S., Klazinga, N. S., and Steyerberg, E. W. Is the readmission rate a valid quality indicator? A review of the evidence. *PloS one*, 9(11):e112282, 2014.

Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Lamb, A., Kukla, M., Pawlowski, N., et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

Glasheen, W. P., Cordier, T., Gumpina, R., Haugh, G., Davis, J., and Renda, A. Charlson comorbidity index: ICD-9 update and ICD-10 translation. *American health & drug benefits*, 12(4):188, 2019.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Huang, K., Altosaar, J., and Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

James, J., Tan, S., Stretton, B., Kovoor, J. G., Gupta, A. K., Gluck, S., Gilbert, T., Sharma, Y., and Bacchi, S. Why do we evaluate 30-day readmissions in general medicine? A historical perspective and contemporary data. *Internal Medicine Journal*, 53(6):1070–1075, 2023.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Karamitros, G., Grant, M. P., and Lamaris, G. A. Associations in medical research can be misleading: A clinician's guide to causal inference. *Journal of Surgical Research*, 310:145–154, 2025.

Kleinberg, S. and Hripcsak, G. A review of causal inference for biomedical informatics. *Journal of biomedical informatics*, 44(6):1102–1112, 2011.

Leppin, A. L., Gionfriddo, M. R., Kessler, M., Brito, J. P., Mair, F. S., Gallacher, K., Wang, Z., Erwin, P. J., Sylvester, T., Boehmer, K., et al. Preventing 30-day hospital readmissions: A systematic review and meta-analysis of randomized trials. *JAMA internal medicine*, 174(7): 1095–1107, 2014.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Mandyam, A., Yoo, E. C., Soules, J., Laudanski, K., and Engelhardt, B. E. COP-E-CAT: cleaning and organization pipeline for ehr computational and analytic tasks. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–9, 2021.

Marafino, B. J., Schuler, A., Liu, V. X., Escobar, G. J., and Baiocchi, M. A causal machine learning framework for predicting preventable hospital readmissions. *arXiv preprint arXiv:2005.14409*, 2020.

McIlvennan, C. K., Eapen, Z. J., and Allen, L. A. Hospital readmissions reduction program. *Circulation*, 131(20): 1796–1803, 2015.

Miley, K., Meyer-Kalos, P., Ma, S., Bond, D. J., Kummerfeld, E., and Vinogradov, S. Causal pathways to social and occupational functioning in the first episode of schizophrenia: uncovering unmet treatment needs. *Psychological medicine*, 53(5):2041–2049, 2023.

Ogarrio, J. M., Spirtes, P., and Ramsey, J. A hybrid causal search algorithm for latent variable models. In *Conference on probabilistic graphical models*, pp. 368–379. PMLR, 2016.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J., Janzing, D., and Scholkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Richens, J. G., Lee, C. M., and Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., and Tsaftaris, S. A. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8), 2022.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Shen, X., Ma, S., Vemuri, P., and Simon, G. Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. *Scientific reports*, 10(1):2975, 2020.

Shi, J. and Norgeot, B. Learning causal effects from observational data in healthcare: A review and summary. *Frontiers in Medicine*, 9:864882, 2022.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000.

Van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., and Forster, A. J. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6):551–557, 2010.

Wang, L., Zhao, L., Luo, Z., Wang, X., Feng, Z., and Lu, L. CARE-30: A causally driven multi-modal model for enhanced 30-day icu readmission predictions. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1509–1516. IEEE, 2023.

Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D. F., Benoit, J., Hughes, M. C., Khan, F., et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1):537, 2021.

Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896, 2008.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International conference on artificial intelligence and statistics*, pp. 3414–3425. Pmlr, 2020.

A. Dataset and Preprocessing Details

Here we provide additional details on cohort construction and preprocessing. We use MIMIC-IV v3.1, a de-identified critical care database containing hospital- and ICU-level EHR data collected between 2008 and 2019. To build the study cohort, patient- and admission-level information are first merged into a base table containing identifiers, timestamps, demographics, and the 30-day readmission and mortality label. We then manually compute the LACE index and its components by combining hospital stay information, emergency department usage, and diagnosis-based comorbidity information. In particular, the comorbidity component is obtained by mapping ICD-9/ICD-10 diagnosis codes to Charlson comorbidity categories and converting the resulting burden score into the LACE-style comorbidity value used in the final score (Glasheen et al., 2019).

To enrich the cohort beyond the handcrafted LACE variables, we use COP-E-CAT (Mandyam et al., 2021) to extract additional variables from the hospital, ICU, and emergency department modules of MIMIC-IV. Since COP-E-CAT is primarily designed for time-series preprocessing, we adapt it to the non-temporal setting considered in this work by collapsing each hospital stay into a single patient-level feature vector. Dynamic variables such as vital signs and laboratory measurements are summarized primarily by their mean value, and in a smaller number of cases by the last observed value, depending on clinical relevance. We then perform additional preprocessing, including variable name harmonization, removal of biologically implausible values, exclusion of variables with high missingness or implausible distributions, and encoding of categorical variables. Finally, to satisfy the i.i.d. assumption required by the causal discovery methods, we retain only the last record per patient and remove redundant admissions. The resulting final dataset contains 56,300 samples and 50 features. The full variable list is provided in Table 2.

A.1. Full Variable List

The final feature set spans demographics, vital signs, laboratory variables, admission details, medication indicators, LACE-related variables, and Charlson comorbidities. Table 2 lists all variables used in the final analysis.

B. Models and Implementation Details

This section summarizes the models used in the study and clarifies how they are applied in this work.

B.1. Predictive Models

We evaluate six tabular prediction models: Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and TabPFNv2. Logistic Regression serves as a simple linear baseline, while Random Forest provides a robust tree-ensemble baseline. XGBoost, LightGBM, and CatBoost are gradient-boosted tree methods that are well-suited to heterogeneous clinical tabular data, and TabPFNv2 is included as a recent tabular foundation model. In the main experiments, we compare these models on two settings: a LACE-only baseline and the full feature set.

Unless otherwise stated, all predictive models are reported with their default hyperparameter settings. In preliminary experiments, we evaluated multiple hyperparameter configurations for each model (e.g., tree depth, learning rate, and ensemble size) and found that performance differences were generally small, while the relative ranking of models remained stable. We therefore report the default configurations for simplicity and reproducibility. For model interpretation, we compute SHAP values on the strongest-performing tree-based predictors. Since these models produce highly similar feature rankings, the main paper reports one representative SHAP visualization. The resulting SHAP-selected variables are then used as fixed inputs to the causal discovery analysis in Section 4.

B.2. Causal Discovery Models

We use three complementary causal discovery approaches. Nonlinear NOTEARS (Zheng et al., 2020) is a score-based method that optimizes a differentiable acyclicity-constrained objective and produces a weighted directed graph. FCI (Ogarrio et al., 2016; Zhang, 2008) is a constraint-based method that relies on conditional independence testing and is designed to tolerate latent confounding, producing a partially oriented graph. DECI (Geffner et al., 2022) is a Bayesian end-to-end causal discovery model that represents uncertainty through a posterior over candidate directed graphs.

We restrict the causal experiments to reduced SHAP-selected feature sets. This reflects the known difficulty of causal structure learning in higher-dimensional settings, where graph recovery becomes less stable, less interpretable, and substantially more

Table 2. Full list of variables.

Category	Variables
Demographics	Age, Gender, PatientWeight
Vitals	HR, NIBP_DIA, NIBP_SYS, NIBP_MAP, RR, SpO2
Laboratory / clinical summary	Physical Status, Hgb, WBC, Glucose, Na, BUN, K
Outcome	Readmitted
LACE-related variables	Length, Acuity, Comorbidity, E, E_origin, LACE
Admission details	Admission Type, Admission Location, Insurance, Marital Status, Race, LOS
Medication flags	Abx_Flag, Diuretic_Flag, Insulin_Flag, Anticoag_Flag, PPI_Flag
Charlson comorbidities	MI, CHF, PVD, CVD, Dementia, COPD, Rheum, PUD, Mild_Liver, DM_no_cc, DM_cc, Paralysis, Renal, Malign, Sev_Liver, Metastatic, HIV

computationally demanding as the number of nodes increases (Schölkopf et al., 2021). For DECI, given that it predicts a posterior distribution over the graphs, we sample 2000 graphs from the posterior distribution and retain only those edges that appear in more than 99% of the sampled graphs. This yields a sparse summary graph that emphasizes highly stable structural relationships while filtering out low-confidence edges.

B.3. Experimental Setup

The predictive and causal pipeline follows a simple sequence:

1. construct the final i.i.d. cohort;
2. train predictive models and evaluate held-out performance;
3. compute SHAP feature importance for the strongest-performing models;
4. select top predictive variables;
5. run causal discovery on the selected feature subsets and compare the recovered structures.

In the main paper, the pooled causal comparison is performed on the fixed top-10 feature subset. The pooled top-5 causal results and the age-stratified robustness analyses are provided in Appendices C and D.

C. Additional Causal Discovery Results

This appendix provides the remaining causal discovery results. The top-5 graphs in Figure 3 provide an aggressive reduction to the most stable core predictors.

Across all three methods, the top-5 graphs preserve the same compact core around readmission, centered on *Physical Status*, *Length*, *LACE*, *Comorbidity*, and *Age*. Relative to the top-10 setting in the main text, the reduced graphs are substantially easier to read but naturally contain less physiological detail. Importantly, the qualitative picture remains unchanged: *Physical Status* remains the most prominent variable in the local neighborhood of *Readmitted*, while *Length*, *LACE*, *Comorbidity*, and *Age* continue to define the surrounding risk structure. This indicates that the main structural conclusions of the paper do not depend on the inclusion of the additional five variables.

The three methods differ mainly in how they connect the variables rather than in which variables are central. NOTEARS provides the clearest weighted structure, FCI yields a more conservative partially oriented graph, and DECI produces the densest connectivity pattern. Despite these differences, all three methods preserve the same small set of structurally important variables. We therefore interpret the top-5 results as additional evidence that the core readmission neighborhood is stable under stronger feature reduction.

D. Age-Stratified Subgroup Analysis

To assess robustness, we rerun NOTEARS after stratifying the cohort into three age groups: 18–44, 45–64, and ≥ 65. We use NOTEARS for this subgroup analysis because it yielded the clearest directed structure in the pooled experiments, facilitating straightforward visual comparison of edge weights across strata. We first show the fixed SHAP-selected top-5 results in Figure 4, which provide the cleanest subgroup comparison, and then show the corresponding top-10 results in Figures 5 to 7.

Across the top-5 subgroup graphs shown in Figure 4, the same core variables remain central across all three age groups:

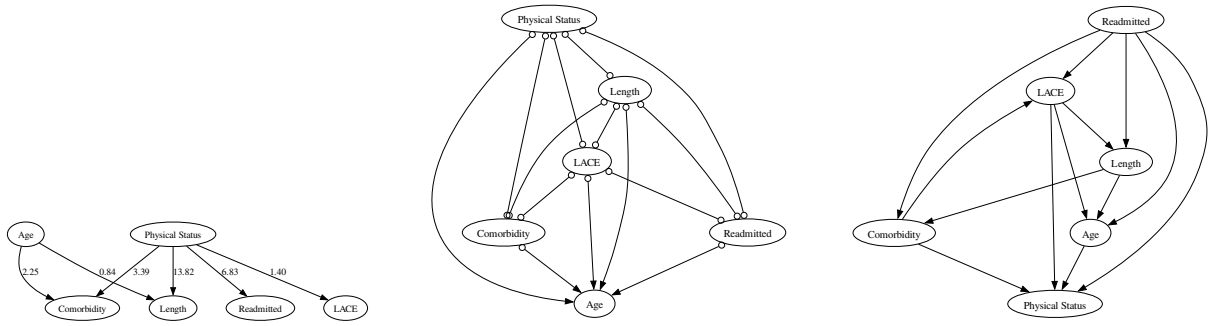

Figure 3. Causal discovery results on the top-5 SHAP-selected features for NOTEARS, FCI, and DECI. All three methods preserve the same core variables, supporting the robustness of the main findings.

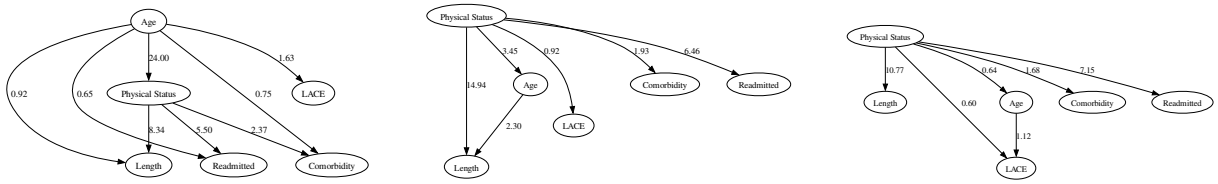

Figure 4. Age-stratified NOTEARS results on the fixed SHAP-selected top-5 feature set for age groups 18–44 (left), 45–64 (middle), and ≥ 65 (right). The subgroup graphs are highly consistent, supporting the robustness of the main structural findings.

Physical Status, *Length*, *LACE*, *Comorbidity*, and *Age*. This is consistent with the pooled top-5 analysis and indicates that the main readmission neighborhood is stable under age stratification. In particular, *Physical Status* remains the dominant upstream node in all age groups, reinforcing its role as the most robust variable across predictive and causal analyses. The subgroup graphs show modest age-dependent differences in edge strength, suggesting that age stratification affects the relative weighting within the recovered structure more than the structure itself.

The top-10 subgroup graphs in Figures 5 to 7 show the same overall pattern while including additional physiological and utilization-related variables. In particular, *Physical Status*, *Length*, *LACE*, *Comorbidity*, and *Age* remain prominent in all groups, while *BUN*, *WBC*, *HR*, *NIBP_MAP*, and *E_origin* provide finer local context. This is consistent with the pooled top-10 NOTEARS result in the main text and suggests that age stratification does not fundamentally alter the set of clinically important variables.

At the same time, the subgroup graphs exhibit modest differences in local connectivity. The younger and middle-aged groups appear somewhat denser in the physiological neighborhood around readmission, whereas the oldest group is more compact and more strongly centered on the core risk variables. This suggests that age stratification changes the relative weighting and local connectivity among variables more than the overall structure. Taken together, the top-5 and top-10 subgroup results support the conclusion that the inferred readmission neighborhood is stable across age strata.

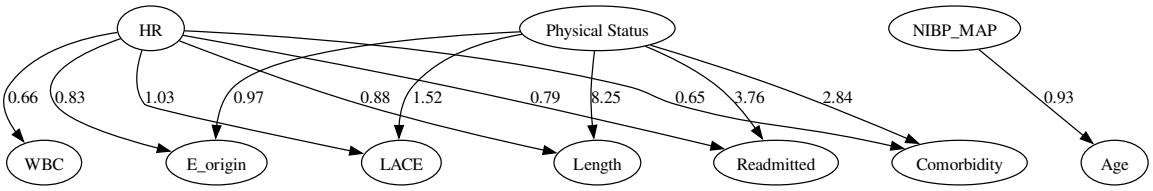

Figure 5. NOTEARS result on the fixed SHAP-selected top-10 feature set for the 18–44 age group.

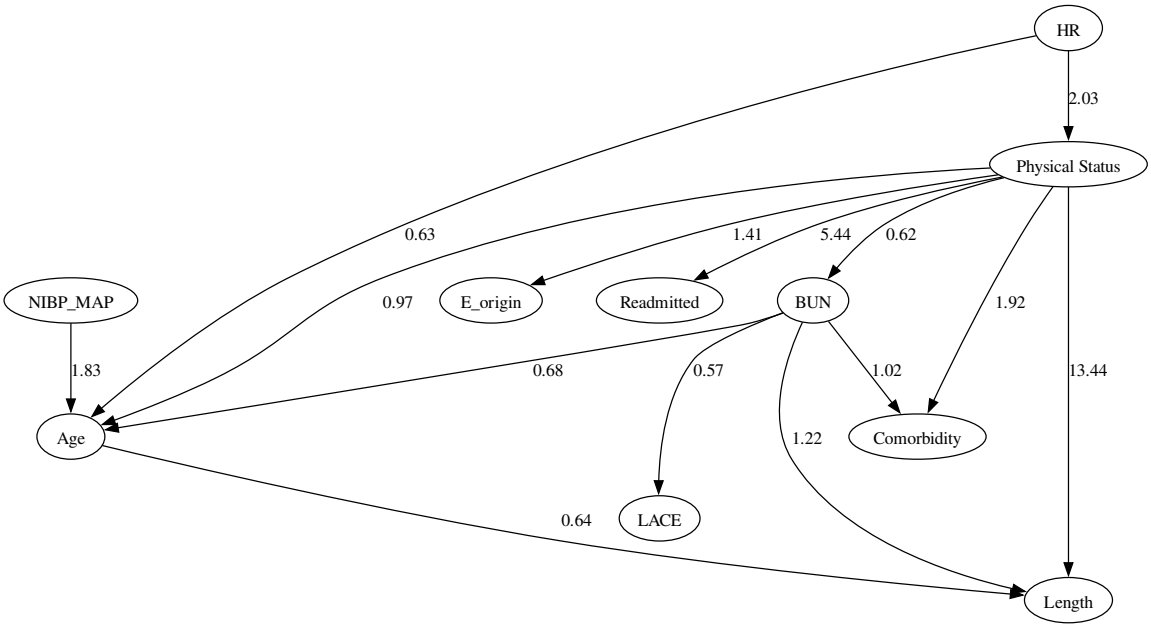

Figure 6. NOTEARS result on the fixed SHAP-selected top-10 feature set for the 45–64 age group.

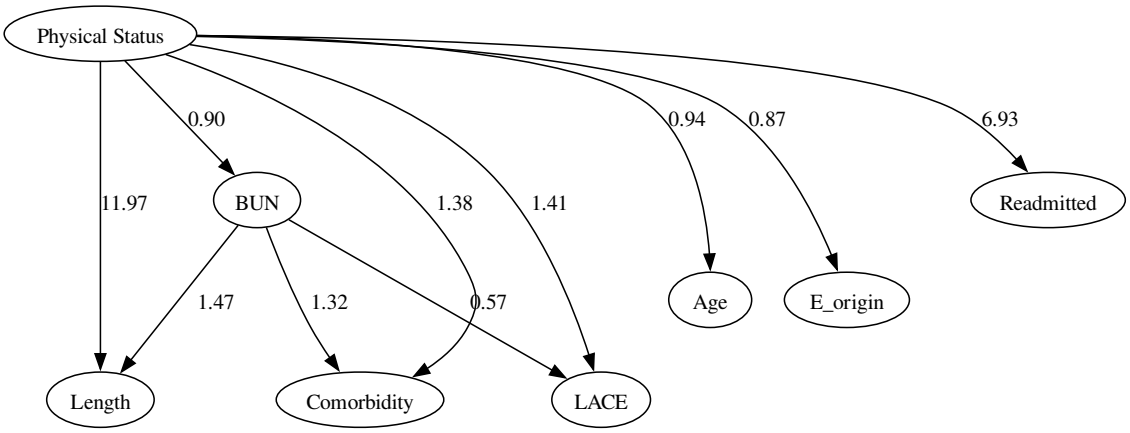

Figure 7. NOTEARS result on the fixed SHAP-selected top-10 feature set for the ≥ 65 age group.