A tutorial on discovering and quantifying the effect of latent causal sources of multimodal EHR data

Marco Barbero-Mota[*]
Department of Biomedical Informatics
Vanderbilt University Medical Center

Eric V. Strobl
Department of Biomedical Informatics
University of Pittsburgh

John M. Still
Department of Biomedical Informatics
Vanderbilt University Medical Center

William W. Stead
Departments of Medicine & Biomedical Informatics
Vanderbilt University Medical Center

Thomas A. Lasko
Departments of Biomedical Informatics & Computer Science
Vanderbilt University Medical Center & Vanderbilt University

Abstract

We provide an accessible description of a peer-reviewed generalizable causal machine learning pipeline to (i) discover latent causal sources of large-scale electronic health records observations, and (ii) quantify the source causal effects on clinical outcomes. We illustrate how imperfect multimodal clinical data can be processed, decomposed into probabilistic independent latent sources, and used to train task-specific causal models from which individual causal effects can be estimated. We summarize the findings of the two real-world applications of the approach to date as a demonstration of its versatility and utility for medical discovery at scale.

1 Introduction

A central objective of biomedical research is to identify the causes of disease. Causal analyses seek to disentangle the effects of interventions and apply this knowledge to guide decisions that improve patient outcomes. However, estimating causal effects for individual patients remains challenging, in part due to latent confounding. Randomized clinical trials address confounding by assigning treatments randomly and therefore remain the gold standard for causal effect estimation. Nevertheless, clinical trials typically investigate only a limited set of interventions (e.g., treatment *vs.* placebo), apply conservative inclusion criteria, and enroll relatively small samples because of their high cost. Consequently, they often provide sufficient data only to estimate aggregate causal effects, such as the average treatment effect, rather than individualized effects conditional on patient covariates, and may suffer from selection bias and limited generalizability [4, 29, 7].

We can attempt to overcome the sample size limitations of clinical trials by approximating randomized experiments with observational data, provided that we condition on the appropriate covariates to mitigate confounding and other biases. The key challenge, however, is identifying which covariates to adjust for. Causal machine learning (ML) models provide a promising framework by enabling proper adjustment even for large covariate sets derived from multimodal Electronic Health Records (EHRs) [9, 8]. Nonetheless, because EHR data are collected primarily for clinical care rather than research, they remain subject to latent confounding, implicit selection biases, and other systemic errors, such as process-related biases, that hinder causal effect

[*]marco.barbero.mota@vanderbilt.edu

1[st]Multimodal Representation Learning for Healthcare Workshop (EurIPS 2025).

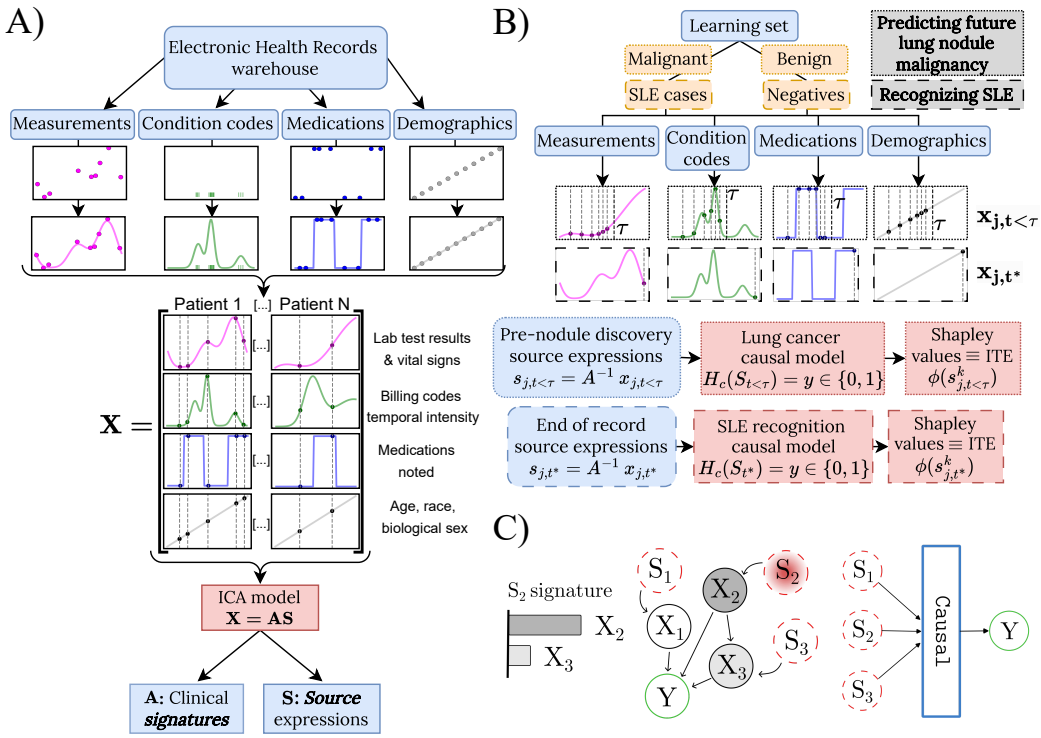

Figure 1: Causal ML pipeline summary. Information flows from multimodal episodic EHR observations that are preprocessed for **A)** independent latent causal sources *discovery*, to **B)** task-specific supervised causal models $H_c(S)$. H_c learns to predict Y from the source expressions of a small labeled cohort. Sample-specific Shapley additive explanations ϕ estimate individual treatment effects at the (j, t) patient-timepoint level for each of $H_c(S)$ k input sources. **C)** The source inputs to $H_c(S)$ are latent exogenous nodes of the underlying linear SCM causal graph. When active, predictive sources leave a *signature* in the observed variables and ultimately affect Y.

identification despite large adjustment sets [34, 1, 21] . Moreover, EHR data introduce additional practical challenges, including systematic and random noise, longitudinal sparsity and incompleteness.

EHR data nevertheless also provide *advantages* that extend beyond the capabilities of randomized trials and other observational datasets. The EHR reflects aggregated practical medical knowledge, captures key aspects of patient physiology, and implicitly records how care is delivered and documented [13]. Clinicians can thus routinely incorporate information from the EHR into their reasoning to infer likely causes of disease compatible with observed clinical states [31], and to evaluate potential interventions most likely to address those causes and improve patient outcomes [32, 9]. We thus ask:

Can we computationally infer latent causes from their effects in the EHR and quantify their individual effects at scale? We argue that the answer is affirmative, though it requires both caution and nuance. Many existing causal ML methods rely on strong untestable assumptions, raising uncertainty about their causal validity when directly applied to large imperfect EHR data [28, 18]. To address this challenge, we created this tutorial to provide a clear, step-by-step summary of the data-driven pipeline introduced in Strobl and Lasko [32] and Strobl [31], and later scaled to large multimodal EHR data in Lasko et al. [22]. We demonstrate that the pipeline (i) addresses many of the intrinsic challenges of processing large EHR datasets; (ii) infers probabilistically independent latent sources that act as candidate causes of EHR patterns in high-dimensional settings; and (iii) quantifies the causal effects of these inferred sources on a patient and timepoint-level for specific outcomes. Note that (ii) and (iii) parallel Steps 1 and 3 of Pearl's counterfactual inference [28]. Crucially, the pipeline circumvents the difficult problem of estimating other intermediary causal structures often required for effect estimation in practice, such as causal graphs or structural equations of the underlying causal processes [28, 32]. We further illustrate its potential through two peer-reviewed real-world applications, framed as predictive tasks in distinct health conditions.

2 Preparing EHR data for machine learning

EHRs provide rich, high-dimensional multimodal data particularly well-suited for ML-based knowledge discovery. However, their episodic nature requires substantial preprocessing to transform sparse, noisy and incomplete observations into a substrate suitable for ML methods. Prior work [20, 19, 22] developed and progressively refined a modular pipeline that can process institution-wide datasets with thousands of variables and a large proportion of missing values into a standardized and complete data matrix ready for downstream analyses. The constituent steps are illustrated in Figure 1 A.

First, we ingest patient visit data for a large unlabeled discovery cohort over four structured data modalities: *measurements*, such as laboratory test results and vital signs; *condition codes* for diagnoses, also referred to as billing codes since their main purpose is to document the complexity of a visit for reimbursement purposes; mentions of *medications*; and *demographics* including race, biological sex and age. Unstructured data modalities such as medical images or clinical notes could also be included by, for example, extracting tabular features from them.

Next, we infer continuous daily-resolution trajectories for each patient variable using modality-specific, clinically-informed methods. This step attempts to overcome within-record data missingness and asynchronicity across clinical variables. *Measurement* curves are smooth piecewise cubic interpolations (PCHIP) [10] that avoid overshooting and maintain monotonicity and nonstationarity where present. *Condition codes* are transformed into intensity curves (i.e. instantaneous code frequency per unit time) using a modified version of Random Average Shifted Histograms (RASH) [6] that accommodates nonstationarity in event-arrival density over a record timespan. *Medication* mentions are converted into binary curves that approximate taking *vs.* not taking by extrapolation from the nearest reconciliation dates. *Demographics* curves are constant one-hot encoded for categorical variables (race and biological sex) and linearly increasing for age. Completely missing *measurements*, *condition codes*, and *medications* in a given record are imputed with population median, a baseline intensity prior of one code per 20 years, and constant 0 (not taking) curves, respectively.

For each patient, the curves are time-aligned into a curveset from which we sample cross sections at a number of random times proportional to the record length and the desired sampling density. Cross sections are then stacked into a dense input matrix X that is then standardized to put all variables on roughly the same scale. For complete data processing details refer to Appendix A in Lasko et al. [22].

3 Discovering causal *sources* and their EHR *signatures*

The matrix X forms the data input to the *discovery step*, with rows representing the EHR variables and columns the record cross-sections. We use the FastICA algorithm to learn an independent component analysis (ICA) model from X. ICA finds a linear unmixing of its input into probabilistically independent latent *sources*. All sources except one are assumed to be non-gaussian. ICA therefore decomposes X into a mixing matrix A and a source matrix S such that $X = AS$ [16, 14]. Rows of S correspond to the mutually independent latent sources, and columns to the cross sections in the corresponding columns of X. The values in S represent the level at which each source is expressed at each cross section. We refer to the columns of A as source *signatures* representing the linear changes in the original data space caused by one unit of expression of a given source. For EHR data, the *i-th* column of A specifies the changes that a one unit change in $S_{i,:}$ imprints to the record across all observed clinical variables.

From Pearl's perspective, a data generation process (DGP) can be formalized through a structural causal model (SCM). Each SCM has an associated directed acyclic graph (Figure 1 C) and a set of equations that encode the functional relationships among exogenous latent sources and endogenous observed variables [28]. Shimizu et al. [30] first described the connection between ICA and linear SCM equations. The identifiability of ICA allows unique recovery of an SCM's exogenous sources, given observations of the endogenous variables, under the linear acyclic non-Gaussian model assumptions [30, 15].

In this context, Strobl and Lasko [32] rigorously defined the sources as *root causes of disease* given observed data. Note that the root level of ICA discovered sources goes as far back in the complete causal graph as the input data allows. For most diseases and EHR datasets, information from the true root causes, such as environmental or genetic factors, is usually upstream from the observed variables in X, and only their effects can be seen in the signatures in A. Each source can reflect pathophysiological pathways, clinical workflow patterns, or a mixture of these [21]. Expert interpretation of a source signature can help identify its clinical meaning and validation, as we will illustrate below.

4 Causally predictive models

Our objective is to find the subset of sources in S that are causally predictive of a health outcome of interest Y. We formulate the problem as a predictive task, where we optimize a supervised architecture to learn to predict Y from source expressions of a labeled patient cohort: $H_c(S) = Y$. Under a weaker unconfoundedness assumption than usual [22, 33], and the assumption that Y is a sink node in the prediction problem causal graph [32], learning from parentless independent nodes constrains the problem such that the only flow of information between the model predictors and Y is through causal paths (Figure 1C). This is in contrast to statistical learning in the original data space ($H_a(X) = Y$), which is susceptible to finding highly predictive but non-causal shortcuts in-sample when minimizing prediction error. In theory, causal identification intrinsic to the sources allows using any supervised learning architecture to learn a descriptive model of Y which confers greater interpretability to clinical ML models.

Given a trained causal model $H_c(S)$, Strobl and Lasko [32] identified sample-specific Shapley additive explanation (SHAP) values [25, 24] as the mathematical formulation for quantifying each source's total causal effect on Y at the instance level; i.e., each source's individual treatment effects (ITE). Although SHAP values are usually employed as an associational measure of feature importance, when computed over the causal model $H_c(S)$, independence among predictors moves up SHAP values from rung 1 in Pearl's ladder of causation, thereby quantifying causal effects [32, 31, 28]. $H_c(S)$ SHAP values estimate each source's attributable impact on Y with respect to a reference "healthy" value, technically referred to as the base value. A source SHAP value captures both its marginal and joint effects from all possible combinations with other sources [32, 17].

5 Real-world examples

Lung cancer [Figure 1C]. Lasko et al. [22] applied the above pipeline to discover 2000 latent causal sources of EHR data for all institution-wide patients with lung disease, defined broadly. From that set of latent causes the authors identified those predictive of *future* malignancy among patients diagnosed with indeterminate pulmonary nodules and no prior history of any cancer. The analysis recovered 92% of established causes of malignant lung cancer. Furthermore, of the top 20 most predictive sources, 6 (30%) were recognized etiologies, 10 (50%) had various level of empirical support in the pulmonary oncology literature, and 4 (20%) had no explicit peer-reviewed evidence but enough clinical face validity to grant further investigation.

Systemic Lupus Erythematosus (SLE) [Figure 1B]. Mota et al. [26] performed the same analysis of EHR latent causes for patients who had an antinuclear antibody test done, presumably under the suspicion of autoimmune disease. A causal model was trained to recognize SLE in the health record, a task also known as computational phenotyping [5]. Predictive sources of SLE chart review labels included multiple disease manifestations: two canonical presentations varying in the specific SLE billing code used, recognizable SLE components (lupus nephritis and four antiphospholipid syndrome forms), a laboratory biomarker of disease activity, SLE treatments, overlapping diseases (rheumatoid arthritis and Sjögren's syndrome), and a surprising operational source capturing a secondary effect of rheumatic autoimmune disorders. Its signature revealed an EHR pattern that clinicians leave when prospectively coding for a rare eye condition, toxic maculopathy, as insurance justification for retinal screening in hydroxychloroquine long-term users. Toxic maculopathy codes are an unconventional predictor of the SLE label, but causal ML was able to identify this process-based latent source.

Results of both studies highlight (i) the value of **causal ML guided by probabilistic independence** as an etiological hypothesis generation at scale, and (ii) its broad applicability across medical specialties. Importantly, both evaluations found no significant in-distribution performance boost for causal models over their correlational counterpart. However, the main objective of causal modeling is not to minimize in-sample predictive error, but rather to find stable predictive descriptors of reality [3], produce higher model interpretability [23], and enable data-driven discovery [11].

Future directions. Currently, our descriptive causal models are trained to minimize the predictive error of $Y = H_c(S)$. However, the loss functions for optimal prediction and optimal inference of internal causal structure are unlikely to coincide. How to choose the combination of hypothesis space, loss function, and optimization approach to maximize causal structure accuracy is an important but unexplored area of research.

A second important and unsolved problem is evaluating causal discovery results, because we lack ground truth for both the identification of latent sources and any counterfactual implications of the model structure [2, 27]. In our work to date, we use domain knowledge to validate inferred causes [22] in an adaptation of quantitative probing, [12] but this does not validate unknown disease causes.

References

[1] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

[2] Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pages 191–201. PMLR, 2019.

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

[4] Amelia J Averitt, Chunhua Weng, Patrick Ryan, and Adler Perotte. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. *NPJ digital medicine*, 3(1):67, 2020.

[5] Juan M Banda, Martin Seneviratne, Tina Hernandez-Boussard, and Nigam H Shah. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annual review of biomedical data science*, 1(1):53–68, 2018.

[6] Mathias Bourel, Ricardo Fraiman, and Badih Ghattas. Random average shifted histograms. *Computational Statistics & Data Analysis*, 79:149–164, 2014.

[7] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

[8] Matthieu Doutreligne, Tristan Struja, Judith Abecassis, Claire Morgand, Leo Anthony Celi, and Gaël Varoquaux. Step-by-step causal analysis of ehrs to ground decision-making. *PLOS Digital Health*, 4(2): e0000721, 2025.

[9] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

[10] Frederick N Fritsch and Judy Butland. A method for constructing local monotone piecewise cubic interpolants. *SIAM journal on scientific and statistical computing*, 5(2):300–304, 1984.

[11] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[12] Daniel Grünbaum, Maike L Stern, and Elmar W Lang. Quantitative probing: Validating causal models with quantitative domain knowledge. *Journal of Causal Inference*, 11(1):20220060, 2023.

[13] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*, 20:117–121, 2013.

[14] Aapo Hyvärinen, Jarmo Hurri, and Patrik O Hoyer. Independent component analysis. In *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, pages 151–175. Springer, 2001.

[15] Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.

[16] Aapo Hyvärinen and Erkki Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(5):411–430, 2000.

[17] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

[18] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

[19] Thomas A Lasko and Diego A Mesa. Computational phenotype discovery via probabilistic independence. In *KDD Workshop on Applied Data Science for Healthcare 2019*, 2019.

[20] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[21] Thomas A Lasko, Eric V Strobl, and William W Stead. Why do probabilistic clinical models fail to transport between sites. *NPJ Digital Medicine*, 7(1):53, 2024.

[22] Thomas A Lasko, William W Stead, John M Still, Thomas Z Li, Michael Kammer, Marco Barbero-Mota, Eric V Strobl, Bennett A Landman, and Fabien Maldonado. Unsupervised discovery of clinical disease signatures using probabilistic independence. *Journal of Biomedical Informatics*, page 104837, 2025.

[23] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[24] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex Degrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. doi: 10.1038/s42256-019-0138-9. URL `https://doi.org/10.1038/s42256-019-0138-9`.

[25] Scott M Lundberg, Paul G Allen, and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 2017. URL `https://github.com/slundberg/shap`.

[26] Marco Barbero Mota, John M Still, Jorge L Gamboa, Eric V Strobl, Charles M Stein, Vivian K Kawai, and Thomas A Lasko. A data-driven approach to discover and quantify systemic lupus erythematosus etiological heterogeneity from electronic health records. In *AMIA Annual Symposium Proceedings*, volume 2024, page 172, 2025.

[27] Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. Validating causal inference methods. In *International conference on machine learning*, pages 17346–17358. PMLR, 2022.

[28] Judea Pearl. *Causality*. Cambridge university press, 2009.

[29] Peter M Rothwell. Factors that can affect the external validity of randomised controlled trials. *PLoS clinical trials*, 1(1):e9, 2006.

[30] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[31] Eric V Strobl. Counterfactual formulation of patient-specific root causes of disease. *Journal of Biomedical Informatics*, 150:104585, 2024.

[32] Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes of disease. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2022.

[33] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[34] Linying Zhang, Yixin Wang, Anna Ostropolets, Jami J Mulgrave, David M Blei, and George Hripcsak. The medical deconfounder: assessing treatment effects with electronic health records. In *Machine Learning for Healthcare Conference*, pages 490–512. PMLR, 2019.