Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)
© 2026 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI260165

Data-Driven and Expert-Informed Causal Discovery for Type 2 Diabetes Risk in Primary Care

Laura AZZIMONTI[a,†], Marta LENATTI[b,1,†], Marco ZAFFALON[a,†], Davide SIMEONE[b,c], Aziz GUERGACHI[d,e], Karim KESHAVJEE[f], Maurizio MONGELLI[b], and Alessia PAGLIALONGA[b]

[a] *IDSIA USI-SUPSI, SUPSI, 6900, Lugano, Switzerland*
[b] *CNR - IEIIT, 10129 Turin, Italy*
[c] *Politecnico di Milano, 20133 Milan, Italy*
[d] *Toronto Metropolitan University, ON M5G 2C3 Toronto, Canada*
[e] *York University, ON M3J 1P3 Toronto, Canada*
[f] *University of Toronto, ON M5T 3M6 Toronto, Canada*

ORCiD ID: Laura Azzimonti https://orcid.org/0000-0003-0565-3505, Marta Lenatti https://orcid.org/0000-0003-0731-7505, Marco Zaffalon https://orcid.org/0000-0001-8908-1502, Davide Simeone https://orcid.org/0009-0004-8510-1421, Karim Keshavjee https://orcid.org/0000-0003-1317-7035, Alessia Paglialonga https://orcid.org/0000-0002-1341-2560

Abstract. The aim of this study is to extract causal relationships from a large set of routinely collected primary care data (biomarkers, medical conditions, risk factors, and medications) from across Canada, with a focus on Type 2 Diabetes risk. The causal discovery process combined data-driven insights with prior expert knowledge, using iterative refinement to construct a causal Directed Acyclic Graph (DAG). The retrieved DAG, which aligns with medical knowledge and performs satisfactorily in predicting future Type 2 Diabetes onset, could serve as a foundation for developing interpretable tools to support medical decision-making.

Keywords. Causal discovery; Electronic Medical Records; Diabetes Prevention

1. Introduction

Developing risk prediction models for preventable chronic diseases such as Type 2 Diabetes mellitus (T2D) can support the implementation of personalized prevention and treatment strategies [1]. Such decision support tools can be developed using causal discovery and inference methods, which extract and quantify cause-effect relationships from real-world data [2]. These methods can be applied to Electronic Medical Records (EMRs) to identify modifiable risk factors that most strongly contribute to the development of a specific disease. Despite their potential to support decision-making, their application in the context of T2D prevention remains limited (e.g., [3-5]).

[†] Equal contribution

[1] Corresponding Author: Marta Lenatti, CNR-IEIIT, Piazza L. da Vinci 32, 20133 Milan, Italy; E-mail: martalenatti@cnr.it.

The aim of this study is to extract potential causal relationships from a large set of routinely collected primary care EMR data, with a focus on T2D risk. The causal graph is learned directly from the data and iteratively refined using expert knowledge to unambiguously determine the orientation of relationships between variables.

2. Methods

2.1. Study dataset

This study used retrospective observational data extracted from the 2000–2015 version of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN, http://cpcssn.ca/), a nation-wide database of primary care EMRs. A waiver of review (REB 2013-261) was granted by the Review Ethics Board of Toronto Metropolitan University as this portion of CPCSSN database included de-identified and anonymized records. The dataset was extracted as follows. Patients under 18 years of age at the time of their first medical encounter and patients with a diagnosis of Type I diabetes or gestational diabetes were excluded, as in [6]. Age, routinely collected blood biomarkers, blood pressure measurements, body mass index (BMI), and medications (cholesterol lowering drugs, corticosteroids, quit-smoking medications, and antidepressants) were extracted for each patient within an extraction window of one year. For patients with T2D (*T2D=1*), this window was defined based on the most recent encounter available up to six months before T2D diagnosis, to compensate for lags between actual T2D onset date and EMR-recorded diagnosis. The median distance between the most recent encounter date and diagnosis date was 1.59 years. For patients with no T2D diagnosis in the database (*T2D=0*), the last available encounter was used to define the data extraction window. The presence of relevant medical conditions (Chronic Obstructive Pulmonary Disease - COPD, hypertension - HTN, osteoarthritis - OA, and depression), was evaluated before the extraction window to preserve temporal ordering between condition onset and observed biomarkers. The most recent information about smoking status (non-smokers, former smokers, current smokers) was considered, when available. Causal discovery was performed on a training set that included 86,618 records with no missing values (T2D=0: 84,340 patients; T2D=1: 2,278 patients). The retrieved causal structure was evaluated on an independent test set of 31,864 records (T2D=0: 30,985 patients; T2D=1: 879 patients). Continuous variables were discretized using thresholds based on clinical guidelines to reflect standard and interpretable risk categories (e.g. for "Pressure": normal pressure, elevated, stage I and stage II hypertension).

2.2. Causal discovery

Causal discovery involves learning causal relationships from observational data, producing a Directed Acyclic Graph (DAG) in which nodes represent observed variables and edges represent inferred causal relationships between them. Specifically, an edge $X \rightarrow Y$ indicates that X is a potential cause of Y. The causal discovery process was conducted by alternating a structure learning step with an expert review step, iterating the process until the resulting DAG was both data-driven and clinically consistent. After each learning step, the estimated DAG was reviewed by a physician with expertise on T2D prediction and prevention (author K.K.) in order to disambiguate the orientation of

edges and remove edges that reflected correlations rather than causal relationships or clinically implausible relationships. The resulting domain expert knowledge was, thus, encoded as blacklists (i.e., sets of forbidden edges in the graph) to constrain the next structure learning step. Specifically:

- ensure that age and sex assigned at birth were treated as root nodes, i.e. non-modifiable factors, by preventing edges from entering these nodes;
- ensure temporal ordering by preventing the T2D node from being the parent of other variables, as the information about T2D was the last available in time;
- ensure that the prescription of a particular group of medications did not affect the prescription of another one;
- ensure that the prescription of a particular group of medications did not lead to the diagnosis of a certain disease, did not cause a change in smoking habits, and did not directly influence the diagnosis of another disease;
- ensure that blood test and pressure values did not influence smoking habits and that they did not influence BMI, assuming that if a causal relationship exists it should be in the opposite direction.

The structure learning step was carried out using score-based estimation methods. Specifically, we maximized the Bayesian Information Criterion through hill-climbing search [7] on the training data, employing bootstrapping to assess the variability of the identified relationships as the input data varied. The bootstrap procedure consisted of 100 iterations, each using 90% of the training set and preserving the original proportion of T2D patients. To evaluate the retrieved causal structure in absence of a ground truth, we assessed its ability to predict T2D onset on the test set. The estimated DAG was complemented with parameters specifying the probability of each node, given its parents. Such parameters were estimated on the training set by means of the Bayesian posterior average under the assumption of uniform prior [7]. We compared the retrieved classification performance with that of a Tree-Augmented naive Bayes (TAN) classifier [7]. To account for class imbalance and maximize sensitivity, the classification threshold was lowered to match the proportion of patients with a future T2D diagnosis in the training set. The analyses were performed in R (v. 4.3.1) using the package *bnlearn* [8]. Main code, final set of blacklists and list of discretization thresholds are available at https://gitlab-core.supsi.ch/dti-idsia/causal_T2D.

3. Results

The estimated causal DAG **(Figure 1)** shows many interconnected pathways between nodes, such as the well-established positive association between age, smoking and COPD, associations between a diagnosis of a given condition and prescription of relatable medications (e.g., from depression to antidepressants, from HTN to antihypertensives) and associations between cholesterol levels (triglycerides - TG, High Density Lipoproteins - HDL, and Low Density Lipoproteins - LDL). Moreover, the probability of T2D onset depends directly on Body Mass Index (BMI) and Fasting Plasma Glucose (FPG), with an indirect contribution from Sex assigned at birth (mediated by BMI) and from Age (mediated by BMI and FPG). The estimated DAG captures the complex interplay between BMI and T2D pathogenesis: elevated BMI may act as a driver for adipose tissue inflammation, β-cells disfunction and insulin resistance, leading to T2D development [9]. In particular, the DAG in **Figure 1** identifies two

distinct pathways that are consistent with those identified in a DAG derived using a retrospective EMR dataset of over 57,000 patients from Mayo Clinic [3].

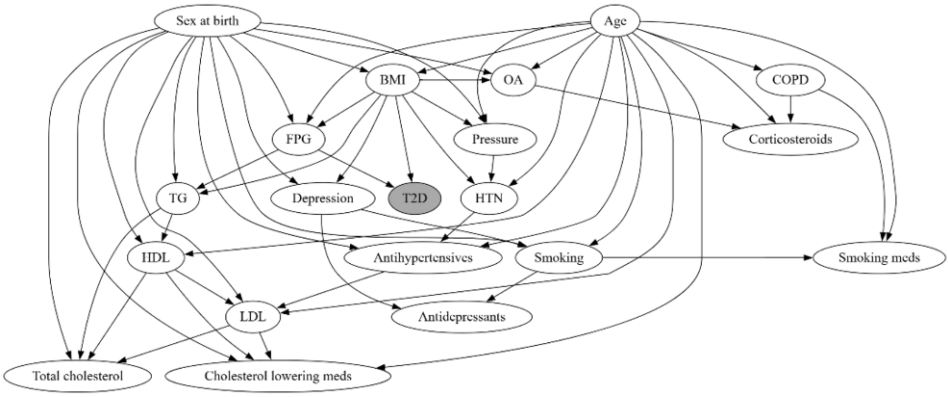


Figure 1. Causal DAG retrieved from CPCSSN data

Stability of the learning process was demonstrated using the bootstrap procedure, which showed 47 out of 51 edges appearing in more than 90% of the iterations. The edge connecting BMI and OA was the only one showing an ambiguous orientation. It was determined a posteriori as "BMI→OA" due to the well-established association between high BMI categories and the incidence of OA, particularly in weight-bearing joints [10]. The predictive model achieved test classification accuracy = 0.75, sensitivity = 0.77, specificity = 0.76, Area under the ROC Curve (AUC) = 0.81, Positive Likelihood ratio (PLR) = 3.16 and Negative Likelihood ratio (NLR) = 0.31 when predicting T2D onset. Performance is comparable to that of the TAN classifier trained on the same data (accuracy = 0.75, sensitivity = 0.74, specificity = 0.75 and AUC = 0.83). Sensitivity, specificity, PLR and NLR values match those of 12 eligible models highlighted in a review on machine learning models for T2D prediction [11] (pooled estimate [95% CI]: sensitivity = 0.81 [0.67-0.90]; specificity = 0.82 [0.74-0.88]; PLR = 4.55 [3.07-6.75]; NLR = 0.23 [0.13-0.42]), whereas AUC results in slightly lower performance with respect to the reported studies (AUC = 0.88 [0.85-0.91]). By focusing solely on Bayesian network models, our model performs better than the highest performing model obtained by [4] on a largely overlapping set of features in terms of AUC (0.76) and sensitivity (0.63), while maintaining comparable accuracy (0.76) and specificity (0.77).

4. Discussion and conclusion

This study examines possible causal interdependencies among a set of commonly available EMR features including blood biomarkers, pressure measurements, diagnoses of chronic diseases and medications, in a large retrospective Canadian cohort captured by a nation-wide database. The resulting DAG is stable across different data samples and aligns with medical knowledge, achieving classification performance comparable to current literature [4,11] while likely representing the main causal relationships underlying T2D onset in the dataset in an interpretable manner. Therefore, it could be used as the starting point to create tools to support disease prevention. In this regard,

future research could focus on using the retrieved DAG to perform counterfactual inference with the aim of predicting the effects of hypothetical interventions (e.g., improvement in diet and physical activity, exemplified through a decrease in BMI and FPG) on T2D risk. This framework may ultimately help clinicians in decision making and support the design of tailored prevention programs by simulating "what-if" scenarios that can quantify the predicted benefits. However, EMR data may not include all factors that significantly influence the observed phenomenon (e.g., poorly coded lifestyle-related information), potentially leading to biased conclusions. To partially address this limitation, future work should explore more sophisticated methods capable of accounting for the effects of unobserved confounders. Moreover, validating the retrieved DAG on a second external cohort will be crucial to ensure the generalizability of the results.

Acknowledgements

This work was supported by the European Union (EU) and by the Swiss State Secretariat for Education, Research and Innovation (SERI) through the project PRAESIIDIUM (GA 101095672). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the EU or SERI. EU and SERI cannot be held responsible for them. M.L. is a PhD student of the National PhD in AI, XXXVIII cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma.

References

[1] Keshavjee K, Candeliere J, Cepeda F, Mittal M, Ali S, Guergachi A. A Framework for Implementing Disease Prevention and Behavior Change Evidence at Scale. Studies in Health Technology and Informatics. IOS Press; 2024 Feb; 312:3-8. doi: 10.3233/SHTI231301.

[2] Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. Nature Medicine. 2024 Apr;30(4):958–68. doi: 10.1038/s41591-024-02902-1.

[3] Shen X, Ma S, Vemuri P, Castro MR, Caraballo PJ, Simon GJ. A novel method for causal structure discovery from EHR data and its application to type-2 diabetes mellitus. Scientific Reports. 2021 Oct;11: 21025. doi: 10.1038/s41598-021-99990-7.

[4] Wang Y, Zhang WS, Hao YT, Jiang CQ, Jin YL, Cheng KK, et al. A Bayesian network model of new-onset diabetes in older Chinese: The Guangzhou biobank cohort study. Frontiers in Endocrinology. 2022 Aug;13. doi: 10.3389/fendo.2022.916851.

[5] Kalia S, Saarela O, O'Neill B, Meaney C, Moineddin R, Sullivan F, et al. Emulating a Target Trial Using Primary-Care Electronic Health Records: Sodium-Glucose Cotransporter 2 Inhibitor Medications and Hemoglobin A1c. American Journal of Epidemiology. 2023 Jan;192(5):782–789. doi: 10.1093/aje/kwad011.

[6] Simeone D, Lenatti M, Lagoa C, Keshavjee K, Guergachi A, Dabbene F, et al. Multi-Input Multi-Output Dynamic Modelling of Type 2 Diabetes Progression. Studies in Health Technology and Informatics. IOS Press, 2023 Oct; 309:228–232. doi: 10.3233/SHTI230784.

[7] Scutari M, Denis JB. Bayesian Networks. Chapman and Hall/CRC; 2021. doi: 10.1201/9780429347436

[8] Scutari M. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software. 2010;35(3):1-22. doi: 10.18637/jss.v035.i03.

[9] Ruze R, Liu T, Zou X, Song J, Chen Y, Xu R, Yin X, Xu Q, "Obesity and type 2 diabetes mellitus: connections in epidemiology, pathogenesis, and treatments," Frontiers in Endocrinology. 2023 Apr;14:1161521. doi: 10.3389/fendo.2023.1161521.

[10] Felson DT, Anderson JJ, Naimark A, Walker AM, Meenan RF. Obesity and Knee Osteoarthritis. Ann Intern Med. 1988 July;109(1):18–24. doi: 10.7326/0003-4819-109-1-18.

[11] Kodama S, Fujihara K, Horikawa C, Kitazawa M, Iwanaga M, Kato K, et al. Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. Journal of Diabetes Investigation. 2022 Jan;13(5):900–8. doi: 10.1111/jdi.13736.