Directed Acyclic Graphs for Identifying Determinants of Postprandial Glucose in Type 1 Diabetes

Rocío Contreras-Jiménez, Juan C. Olivares-Rojas, Adriana C. Téllez-Anguiano, José A. Gutiérrez-Gnecchi, J. Eduardo Alcaráz-Chávez, Enrique Reyes-Archundia

Tecnológico Nacional de México/Instituto Tecnológico de Morelia
Morelia, México
rocio.cj@morelia.tecnm.mx

Abstract— **Type 1 Diabetes Mellitus (T1DM) requires precise glucose management to prevent complications. A significant challenge is determining the optimal insulin bolus dose, as errors can lead to hypo- or hyperglycemia. This paper presents an application of Directed Acyclic Graphs (DAGs) to model causal relationships that influence postprandial glucose levels using the HUPA-UCM dataset (25 T1DM patients with glucose, insulin, activity, and dietary data). Three DAG models were generated: knowledge-based (DAG1), data-driven (DAG2), and adjusted (DAG3). Causal inference was performed using DoWhy, revealing a statistically significant effect of insulin bolus on postprandial glucose. Each additional bolus unit reduced glucose between −1.30 mg/dL (DAG1) and −0.78 mg/dL (DAG2/DAG3). Robust tests confirmed stability ($p > 0.90$). Results demonstrate that DAGs can reliably identify determinants of glycemic response and represent a promising tool for interpretable, personalized insulin dosing systems. These results could support the design of intelligent insulin dosing systems capable of explaining their recommendations to clinicians and patients.**

Keywords— Type 1 Diabetes, Health technology, causal inference**, Directed Acyclic Graphs, Insulin Bolus calculation**

I. Introduction

Type 1 Diabetes Mellitus (T1DM) is a chronic, incurable autoimmune disease characterized by the destruction of pancreatic beta cells, which prevents the production of endogenous insulin [1]. Patients require lifelong exogenous insulin administration, and maintaining glucose levels between 70 and 180 mg/dL is critical to avoid acute and chronic complications. Incorrect insulin dosing may result in hypoglycemia, hyperglycemia, diabetic ketoacidosis, or long-term damage such as nephropathy, retinopathy, or cardiovascular disease [2], [3], [4].

An essential justification for doing this work is that the Glycemic imbalance in patients with T1DM, whether hypoglycemia or hyperglycemia, can cause short, medium and long-term effects, which can affect not only the quality of life of patients and family caregivers, but also high treatment costs not only for the patient and their family but for the public health system, also affecting the patient's work productivity and can decrease or eliminate their economically active life, which is why it is essential to develop tools that allow patients and their families to maintain adequate glycemic control, which enables the patient to have a better quality of life in the short and long term, reducing the costs of specialized treatments, operations and hospitalizations and even premature death. Better glycemic control could improve patients' quality of life, reduce the risk of complications above, and decrease the financial burden on patients, their families, and government health systems. According to the National Institute of Statistics, Geography, and Informatics (INEGI), diabetes is the second leading cause of death in Mexico, with 84,095 deaths in 2024, according to the report published in 2025 [5], surpassed only by heart disease. Regarding costs, the Mexican Diabetes Federation reported in 2019 that the annual costs for a controlled patient were approximately 88,024 Mexican pesos, while those for an uncontrolled patient could reach 1,163,028 Mexican pesos. This expense not only affects the economy of patients and their families, but also that of the public health service [6]. Currently, some insurers confirm that the treatment costs of an uncontrolled patient rise to several thousand pesos [7] and the Mexican Social Security Institute (IMSS) declared in its report to the Federal Executive and the Congress of the Union that Diabetes is the most expensive disease for the IMSS, with a medical care cost of 106 million pesos per day [8].

The main engineering challenge lies in calculating the optimal rapid-acting insulin bolus to cover meals, a task traditionally performed with the carbohydrate formula and correction factors. Recent advances include continuous glucose monitoring (CGM), insulin pumps, and machine learning (ML) algorithms. However, most approaches rely on associations and correlations, without addressing the causal relationships between physiological and behavioral variables. This limits their interpretability and generalizability.

However, most ML approaches are correlation-based, limiting interpretability. Causal inference, particularly Directed Acyclic Graphs (DAGs), enables explicit modeling of cause-and-effect relationships, allowing for intervention and counterfactual reasoning, which are essential to clinical decision-making.

This work applies DAGs to analyze insulin bolus effects in T1DM using the HUPA-UCM dataset [9]. Three models were constructed—knowledge-based, data-driven, and adjusted—to estimate the causal impact of bolus insulin on postprandial glucose. Findings support the potential of DAGs to complement conventional algorithms and improve individualized insulin dosing strategies.

979-8-3315-9026-0/25/$31.00 ©2025 IEEE **Online ISSN: 2332-5712**

The HUPA-UCM dataset [9] was selected for this study because it contains high-quality continuous glucose monitoring and lifestyle data from 25 patients with T1DM, including glucose levels, insulin doses, carbohydrate intake, steps, calories, and basal rates. Although relatively small, this dataset integrates relevant factors that are often overlooked in traditional bolus calculators.

Derived from the challenge of maintaining ideal glycemic control in a patient with T1DM, allowing for a glucose level between 70 and 180 mg/dL without the risk of hyperglycemia and hypoglycemia, it is necessary to be able to calculate an insulin bolus as accurately as possible. To achieve this goal, various research efforts have been undertaken, ranging from the application of carbohydrate-based insulin bolus calculation formulas and glucose correction doses [4] to the use of insulin pumps with continuous glucose monitors, smart pens, and Artificial Intelligence-based applications. However, in a country like Mexico, these advances are not available to all patients. In this article, the use of Directed Acyclic Graphs is proposed, evaluating not only the pre-meal glucose (before the three main meals of the day) and the amount of carbohydrates to be consumed, such as the carbohydrate-based formula and some of the existing calculators, but also considering other factors that influence the patient's glucose, such as: date, time, basal rate, pre-meal glucose, calories consumed, carbohydrates, steps, and insulin bolus applied, this derived from the variables contained in the dataset that was used, which is the HUPA-UCM [9], on the other hand, the calculation and search for post-meal glucose (two hours after the meal) was made, in the glucoses of the same dataset, with these variables, the causal model of Directed Acyclic Graphs was constructed, to find a causal inference that tells us which variables and to what extent they influence the patient's post-meal glucose.

II. Related Works

To date, various applications and algorithms have been developed to calculate insulin boluses, using different monitoring variables and glucose measurement methods, including traditional glucometers, CGMs, and closed-loop systems in conjunction with insulin pumps. Cappon et al work with continuous glucose monitors, using the Uva-Padova simulator dataset and a noise-free single-meal scenario with varying conditions, including the amount of food, pre-prandial glycemia, and Glucose Rate of Change values [10]. The Random Forest and Gradient Boosting Trees model is a non-linear model that integrates variables such as the glucose rate of change, using their own dataset. They utilize IoT devices to acquire real-time blood glucose data from the CGM sensor and an insulin pump. The variables of the dataset are patient ID, Age, Sex, Body Mass Index, Glucose before food, bolus insulin administered, date and time of the bolus, postprandial Glucose, Carbohydrate ratio, Correction factor, and Basal Infusion [11].

There are deep learning models that integrate layers of long-term memory networks and gated recurrent units to estimate basal insulin and insulin bolus. Most algorithms perform insulin bolus calculations using the standard formula, which includes the following variables: CHO (g), representing the carbohydrate intake of the meal; the insulin-to-carbohydrate ratio; the correction factor; and a few additional variables. Recent studies have emphasized the relevance of causal inference in healthcare for robust, explainable decision-making [12] [13]. After reviewing the state of the art, it was decided to select DAGs because they can explicitly represent causal relations among clinical and lifestyle variables. In contrast, traditional statistical models can only identify associations. DAG's facility for inferring causality from correlation reduces the risk of bias from confounding variables. Causal inference encompasses a range of models and several algorithms. This article will focus on the DAG (Directed Acyclic Graph) model, a discrete mathematical structure that facilitates inference and causal discovery to understand the relationships among a larger number of variables.

III. Theoretical Framework

For a better understanding of the content of this article, it is necessary first to explain some concepts.

A. Causal Models

"A causal model is an abstract quantitative representation of real-world dynamics. Therefore, a causal model attempts to describe the causal and other relationships between a set of variables" [14]. One of the reasons why causal inference was chosen to obtain a more accurate insulin bolus calculation is that the association and correlation of variables that has been applied so far in most existing algorithms, even those using ML, is that causality allows us to make interventions and counterfactuals, which could help us detect causal inference in observational data without putting the health of a real patient at risk.

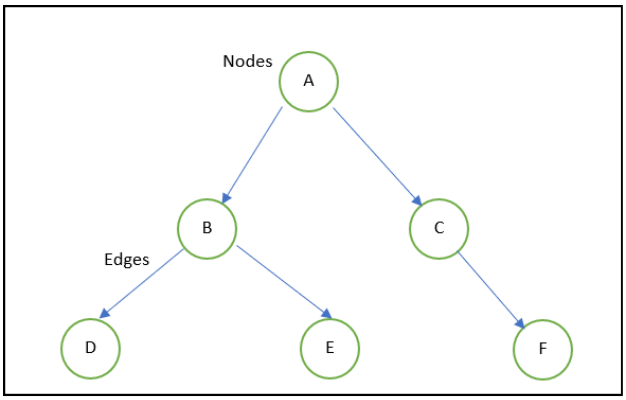
The concept of the ladder of causality, introduced by Judea Pearl has three rungs: the first, which is the association is related to observation, in such a way that we learn by observing for example historical data and we can realize how observing one thing can change our belief about another, the second rung is the intervention, where we do something to intervene in the result of what was previously observed, many of the existing algorithms work up to this level, with the prediction, by intervening for example the amount of insulin bolus, or the amount of carbohydrates and predicting the result, however the third rung of the ladder of causality, which represents the counterfactuals, represents imagining and understanding, which would be asking ourselves if we had done something, if this would change the result, when dealing with health issues in human beings, this last rung undoubtedly represents a possibility that can hold great hopes for progress without putting the health of real patients at risk [15], and [16]. Recent studies have emphasized the relevance of causal inference in healthcare for robust, explainable decision-making.

B. DAG

"A directed acyclic graph (DAG) is a type of graph in which nodes are linked by unidirectional connections that do not form any cycle. DAGs are used to illustrate dependencies and causal relationships." [17], are made up of nodes (circles) that can represent entities, objects or in our case variables that influence others or our final result which would be postprandial glucose (glucose two hours after consuming food) and arrows or edges that represent a unidirectional relationship, where the way in which some nodes affect others is graphically represented and

are connections between variables or entities, which when directed, are traveled in only one direction and represent that a node has a causal relationship on another, as shown in Figure 1.

Fig. 1. Example of a directed acyclic graph



C. Medical Applications

The application of causal inference has been identified as a significant area of opportunity in clinical medicine and public health, especially "in observational studies to quantify the effect of an exposure or intervention on an outcome (causal inference)" [13], given that randomized clinical trials are the best option to evaluate interventions in the health area, but many times they cannot be carried out, due to the risk that would be put to patients or for ethical reasons, however, decision makers, patients, caregivers, and doctors or health personnel, must make these decisions, due to the above, observational studies based on simulators, application of Artificial Intelligence and causal inference become necessary [13]. Although the application of AI is something new, the application of causal inference in medicine was already used as "part of everyday life, since man has always sought the why of things, as a way of understanding and adapting to the world. In medicine, one of the central objectives of study is the identification of the factors or agents that cause diseases, to establish treatments and apply preventive measures" [12].

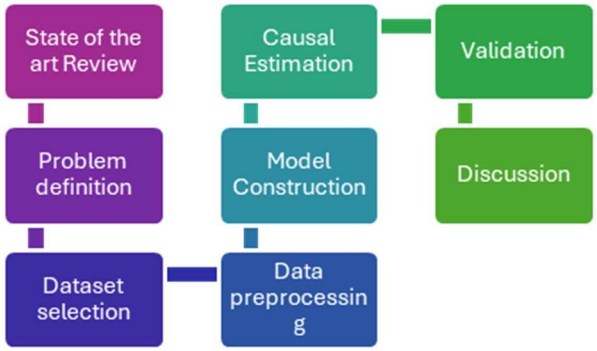
IV. Methodology

The proposed methodology for applying Directed Acyclic Graphs to generate a causal model that explains the causes influencing postprandial glucose consisted of the following steps, as shown in Figure 2:

All analyses were performed using the DoWhy (v0.10) and NetworkX (v3.3) libraries in Python with Google Colab.

1. State of the art review. Identification of previous ML models for insulin bolus calculation.
2. Problem definition. Treatment: bolus_volume_delivered; Outcome: glucose_postprandrial.
3. Dataset selection. HUPA-UCM (25 T1DM patients, CGM + lifestyle data). Different datasets with T1DM information were searched, and the HUPA-UCM diabetes dataset was selected for further analysis. “This dataset provides a collection of CGM data, insulin dose administration, meal intake counted in grams of carbohydrate, steps, calories burned, heart rate, and sleep quality and quantity assessment acquired from 25 individuals with type 1 diabetes mellitus (T1DM). CGM data were acquired using FreeStyle Libre 2 CGMs, and Fitbit Ionic smartwatches were used to collect data on steps, calories, and heart rate for at least 14 days. This dataset could be used directly from the preprocessed version or customized from raw data.” [9]. There are 28 files, one for each patient, containing the following preprocessed data: time, glucose, calories, heart rate, steps, basal rate, bolus volume delivered, and carb input
4. Data preprocessing. Cleaning, calculation of postprandial glucose (2h after meal), normalization, and categorical encoding.
5. Model construction. DAG1: knowledge-based., DAG2: data-driven., DAG3: knowledge + adjustment.
6. Causal estimation. Backdoor and instrumental variable adjustments using DoWhy.
7. Validation. Refutation tests with random confounders.
8. Discussion. Interpretation and comparison of models.

Fig. 2. Diagram of the proposed methodology



V. Results

A. Phase 1: Problem Configuration:

First, the graph representing the problem was created, and an instance of that object was created. The Dowhy inference and causal knowledge library was used to generate three DAGs: the first based on prior knowledge, the second based on data, and the third, a model based on previous experience but with adjustment.

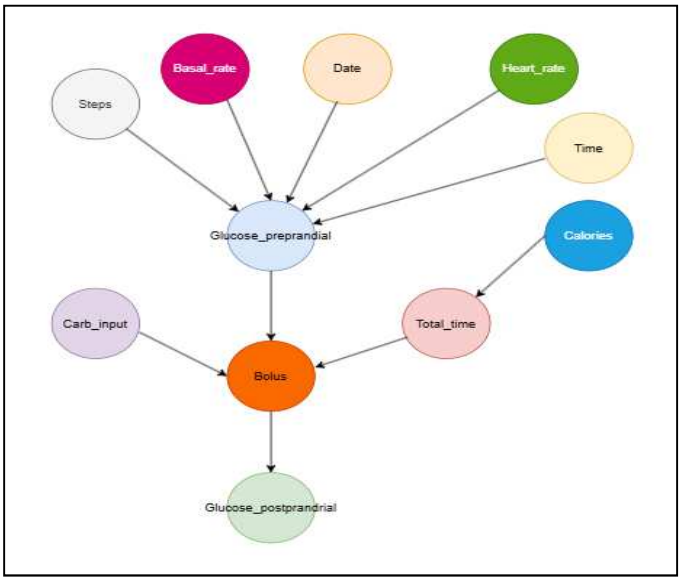
DAG1: Model based on prior knowledge, documentary research, and experience. Direct effect of bolus on postprandial glucose, conditioned only on preprandial glucose. Causal relationships were established, and the model is presented in Figure 3. In this Model, one can see only the variables and the direction of the arrows, not the weights of each variable; however, it shows the dependencies.

DAG1, based on prior clinical knowledge, assumes that preprandial glucose levels and lifestyle factors influence both bolus dose and postprandial glucose levels.

DAG2: The graph was constructed from the HUPA-UCM dataset, with bolus_volume_delivered as the treatment or variable to be tracked, and glucose_postpandrial as the output

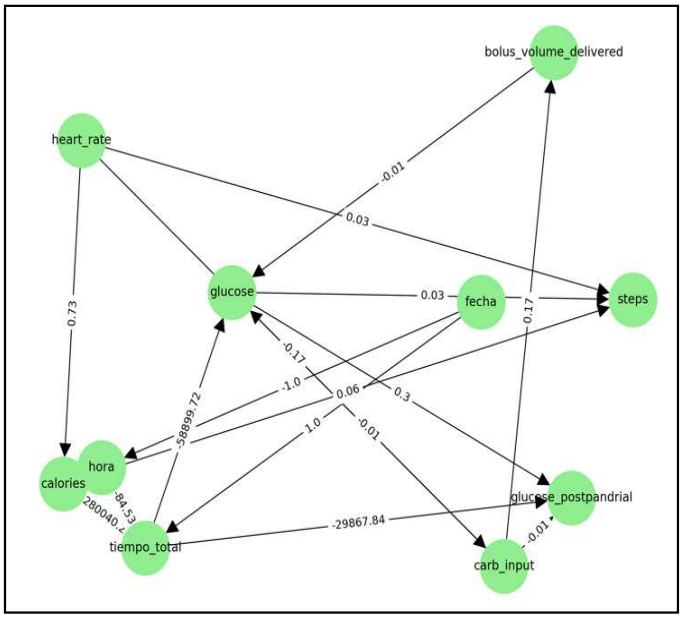
variable believed to be influenced by it, resulting in the graph shown in Figure 4.

Fig. 3. Directed Acyclic Graph based on prior knowledge DAG1



In DAG2, it is extracted from relationships directly detected in the data. Most variables are directly related to postprandial blood glucose levels. The basal_rate variable does not appear in this graph, as the Python program used to eliminate variables that tend toward zero. This directed acyclic graph has the weights for each variable, enabling us to understand how they influence postprandial glucose levels.

Fig. 4. Directed Acyclic Graph based on data DAG2.



B. Phase 2: Identify the causal parameter or parameters.

In DAG3, shown in Figure 5, you can see how an adjustment is made to DAG2 based on knowledge.

In Table I, the results of the Estimand calculations of the three generated models are summarized.

Fig. 5. Directed Acyclic Graph based on prior knowledge with adjustment DAG3.

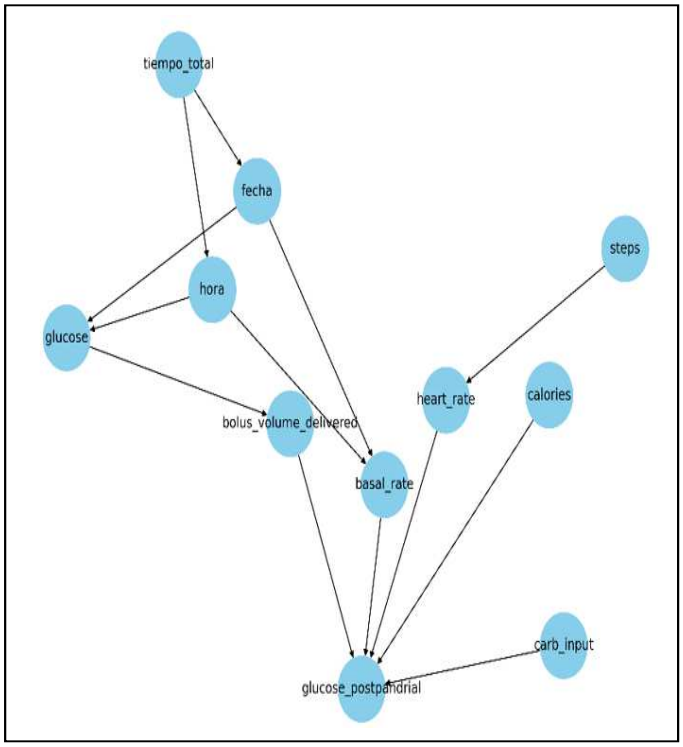


TABLE I. CAUSAL PARAMETERS IDENTIFIED IN DAGS

Model	Principal Estimand	Adjustment type	Conclusion
DAG1 Prior knowledge	Backdoor: Effect of bolus on postprandial glucose, depending on glucose. Instrumental variable: carb_input and calories	Backdoor and Instrumental Variable (IV)	It allows estimating the causal effect by adjusting for pre-pandemic glucose and IV use.
DAG2 (Data Driven)	Backdoor: Adjustment for multiple covariates (e.g., glucose, calories). Estimate 4: (general_adjustment) matches backdoor.	Backdoor	Estimation by adjustment for a set of variables, without IV or frontdoor.
DAG3 Prior knowledge with adjustment	Backdoor: Conditional adjustment in glucose (prepandrial).	Backdoor	A valid estimate requires that glucose adequately represent insulin requirements; additional covariates could be used.

C. Phase 3: Get the estimates

Model DAG1 has a Method Backdoor adjustment (single variable: glucose). Estimator: Nonparametric ATE. The effect of the insulin bolus is estimated conditional only on the prior glucose level. Result: Estimated causal effect: −1.2994 and p-value: 0.96. It means a statistically significant effect. Each additional bolus unit reduces postprandial glucose by an average of 1.3 units. Without adjusting for other variables, the estimated impact is substantial, but it may be overestimated because other potential confounders were not considered.

DAG2 Model has the Method Backdoor with Adjustment (multivariate). The Adjustment variables are heart rate, total time, basal rate, glucose, calories, steps, time, date, and carb input. Estimator: Nonparametric ATE. The Result: Estimated causal effect: −0.786 and p-value: 0.96. It has a statistically significant effect when controlling for a broader range of factors (activity, carbohydrates, basal, etc.). The bolus effect diminishes, indicating a more precise and less biased estimate than in DAG1.

Model DAG3 has the same Method as DAG2 (same adjustment set), with an estimator of Nonparametric ATE. The Result shows an estimated causal effect of −1.768 and a p-value of 0.90. It corroborates the estimate obtained in DAG2. The model is more robust and reliable by incorporating the whole clinical context while maintaining statistical significance.

D. Phase 4: Perform disconfirmatory tests to validate the models

To assess the stability of the causal estimates obtained in models DAG1, DAG2, and DAG3, the "Add a random common cause" refutation test was applied. This test involves introducing a random variable that acts as a spurious confounder and observing whether the estimate of the causal effect changes significantly. The results are presented in Table II below.

TABLE II. ANALYSIS OF REFUTATION TESTS IN MODELS DAG1, DAG2, AND DAG3

Model	Estimated effect	Effect with random confounder	Difference	Value p
DAG1 Prior knowledge	-1.2994	-1.2994	~0.00001	0.96
DAG2 (Data Driven)	-0.7860	-0.7861	~0.00006	0.92
DAG3 Prior knowledge with adjustment	-1.7685	-1.7684	~0.00015	0.90

In all three models, the p-value was greater than 0.90, indicating that the difference between the original estimated effect and the new effect after the inclusion of a random confounder is not statistically significant. Furthermore, the absolute change in the estimate was minimal across all models, at less than 0.0002 units.

According to the DAG1 model and the assumption that conditioning on prior glucose eliminates all confounding, an increase in the insulin bolus results in a statistically significant decrease in postprandial glucose, with an average effect of -1.3 units.

In the DAG2 model, when controlling for a broader set of variables (including physical activity, food intake, and time of day), each unit of insulin bolus is estimated to reduce postprandial glucose by 0.78 units, a statistically significant decrease.

In the DAG3 model, the results indicate that, considering multiple physiological and contextual factors, the insulin bolus has a significant negative impact on postprandial glucose levels. That is, it consistently reduces glucose after eating.

Compared to the results of the DAG1 model, where the estimated effect was more substantial (around -1.30), the effect here is more moderate. This suggests that the DAG3 model reduces bias by adjusting for more relevant variables. The estimated effect is more dependable because it controls activity, diet, and circadian rhythm, in addition to prior glucose levels.

VI. DISCUSSION OF THE RESULTS

Three DAG models were generated. DAG1, based on clinical knowledge, assumes preprandial glucose and lifestyle factors influence bolus dose and postprandial glucose. DAG2 was data-driven, while DAG3 integrated prior knowledge with adjustments. The estimated causal effects (Nonparametric ATE) were: −1.30 mg/dL (DAG1), −0.786 mg/dL (DAG2), and −1.786 mg/dL (DAG3). All models were statistically significant ($p < 0.01$). 95% confidence intervals ranged from −0.95 to −0.62 mg/dL.

Refutation tests adding random confounders yielded negligible effect changes ($\Delta < 0.0002$, $p > 0.90$), confirming robustness. Thus, the results are internally valid and stable across model variations.

These results will support the design of an intelligent insulin dosing system capable of explaining its recommendations to clinicians and patients.

A. Limitations

The main limitations include the small sample size (25 patients), potential selection bias from the dataset, and the assumption of non-linear relationships approximated by nonparametric estimators. Moreover, some variables, such as basal rate or steps, that do not reflect exercise intensity may introduce noise into the measurements, potentially affecting estimation precision. Although the dataset comprises only 25 patients, its multimodal nature (CGM, insulin, activity, and nutrition) provides high temporal resolution suitable for causal analysis. The project will run the models with other datasets and, in the near future, aims to generate its own dataset using real data from volunteer patients.

VII. CONCLUSION

The results demonstrate the robustness of the causal estimates against the presence of spurious variables. This suggests that the effects of insulin bolus volume on postprandial glucose levels, as estimated by the DAG1, DAG2, and DAG3 models,

are reliable and not due to spurious associations artificially introduced. Consequently, the internal validity of the causal analysis developed in this research could inform the next step and enable predictions. Currently, our work is applying additional causal algorithms and generating synthetic data to allow the inclusion of a larger number of records. We are also working on transformations that enable data analysis using models that assume normal distributions and linear relationships. This study aims to identify models that can achieve greater efficiency in calculating insulin boluses and better explain patients' physical behavior, thereby improving their quality of life in the short, medium, and long term. DAG-based causal modeling proved effective for identifying determinants of postprandial glucose in T1DM. The consistent, statistically significant negative effect of an insulin bolus on glucose levels demonstrates the utility of causal inference in health technology. DAGs bridge clinical knowledge with data science, enabling explainable and personalized decision support systems.

VIII. Future Work

Future research will explore hybrid DAG–machine learning architectures, the use of other causal inference algorithms, and the integration of causal reasoning with deep learning models for personalized insulin dosing. Expanding the dataset and using other datasets are also planned.

Acknowledgments

The authors thank the National Technological Institute of Mexico for supporting this project with the grant 21636.25-P.

References

[1] World Health Organization, "Diabetes," 14 11 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed 16 07 2025].

[2] D. J. P. Hayes, "«Diabetes Mellitus Tipo 1»," vol. 06, 2008.

[3] Federación Mexicana de Diabetes, "Qué es la insulina?," Federación Mexicana de Diabetes, 2022. [Online]. Available: https://fmdiabetes.org/que-es-la-insulina/.. [Accessed 17 07 2025].

[4] Centro de enseñanza sobre la Diabetes de la Universidad de California en San Francisco, "Cálculo de la dosis de insulina," Universidad de California en San Francisco, [Online]. Available: https://diabetesteachingcenter.ucsf.edu/about-diabetes/type-2-diabetes/use-insulin-type-2-diabetes/calculating-insulin-dose. [Accessed 16 07 2025].

[5] INEGI, "ESTADÍSTICAS DE DEFUNCIONES REGISTRADAS (EDR)," 25 02 2025. [Online]. Available: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2025/edr/EDR_En-sep2024.pdf. [Accessed 16 07 2025].

[6] Federación Mexicana de Diabetes, "Los costos de la diabetes," 09 01 2019. [Online]. Available: https://fmdiabetes.org/los-costos-la-diabetes/. [Accessed 15 01 2023].

[7] Metlife, "¿Cuánto cuesta vivir con diabetes en México?," 01 11 2024. [Online]. Available: https://www.metlife.com.mx/blog/bienestar-financiero/gastos-de-vivir-con-diabetes-cuanto-necesitas/. [Accessed 16 07 2025].

[8] S. Díaz Mora, "Diabetes, la enfermedad más costosa para el IMSS; su atención médica cuesta 106 millones de pesos al día," *El economista,* 12 07 2025.

[9] A. J. B. M. A. A. V. J. M. G. O. Hidalgo J. I., *HUPA-UCM diabetes dataset.,* Madrid: Mendeley Data, 2024.

[10] M. V. F. M. A. F. G. S. Giacomo Cappon, "A Neural-Network-Based Approach to Personalize Insulin Bolus Calculation Using Continuous Glucose Monitoring," *Journal of Diabetes Science and Technology,* vol. 12, pp. 265-272, 2018.

[11] L. Z. y. S. Zhong, "«A Health Management Platform based on CGM Device»," *IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) ,* pp. 822-826, 2019.

[12] H. &. P.-C. E. Álvarez-Martínez, "Causalidad en medicina.," *Gaceta médica de México ,* pp. 467-472. , 2004.

[13] I. Nunez and M. Lajous, "El ensayo diana para la inferencia causal en estudios observacionales," *Salud pública de México,* vol. 67, no. 1, pp. 83-90, 2025.

[14] Oxford University, "Oxford Reference," 2025. [Online]. Available: https://www.oxfordreference.com/display/10.1093/oi/authority.20111005143410337. [Accessed 18 07 2025].

[15] J. Pearl and Mackenzie, The book of why, Penguins Books, 2019.

[16] A. Molak, Inferencia y descubrimiento causal en Python, Madrid: Ediciones Anaya Multimedia (Grupo anaya S. A.), 2024.

[17] A. Gomstyn and A. Jonker, "¿Qué es un gráfico acíclico dirigido (DAG)?," 06 03 2025. [Online]. Available: https://www.ibm.com/mx-es/think/topics/directed-acyclic-graph. [Accessed 18 07 2025].