Local causal discovery in epidemiology: an application to quantifying the effect of diabetes on severe liver fibrosis in patients with viral hepatitis

Timothée Loranchet[1], Daria Bystrova[1,2], Paul Burgat[1], Jonathan Bellet[1], Marc Bourlière[3,4], Clovis Lusivika-Nzinga[1], Jerome Nicol[1], Lucia Parlati[5], Pierre-Yves Boëlle[1], Fabrice Carrat[1,6], and Charles K. Assaad[1]

[1]Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, Paris, France
[2]France Cohortes, INSERM, Paris, France
[3]Département d'hépatologie et gastroentérologie, Hôpital Saint Joseph, Marseille, France
[4]Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France
[5]Université de Paris Cité, INSERM, AP-HP, Département d'Hépatologie/Addictologie, Hôpital Cochin, Paris, France
[6]Hôpital Saint-Antoine, Unité de Santé Publique, AP-HP, Paris, France

Abstract

Background Assessing the controlled direct effect (CDE) of an exposure on a disease from observational data is challenging when the DAG is unknown. In such cases, causal discovery methods can be used to infer a partiallly oriented DAG from which potential adjustment sets for effect estimation can be deduced. However, classical causal discovery approaches that attempt to reconstruct the entire graph often face important limitations, including strong assumptions, numerous statistical tests, and large sample size requirements, which are further exacerbated by missing data. To address these issues, we rely on a local causal discovery algorithm, which focuses on identifying only the relevant part of the graph. We apply this approach to cohort data from patients with viral hepatitis to estimate the CDE of diabetes on severe liver fibrosis.
Methods We investigated the CDE of diabetes on liver fibrosis among patients infected with HBV or HCV using baseline data from the French ANRS CO22 HEPATHER cohort. We used a local causal discovery algorithm, called, LocalPC-CDE. This method focuses on learning only the part of the graph necessary to answer the target causal question, thereby avoiding some unrealistic assumptions, reducing the number of required tests, and automatically selecting the adjustment set for valid CDE estimation. To further enhance robustness, we incorporated a bootstrap augmentation of the algorithm to quantify the stability of each identified direct cause (variables of the adjustment set). Only the variables least affected by sampling fluctuations were retained in the final adjustment set. When the CDE was identifiable, it was estimated using logistic regression, yielding odds ratios (OR) with 95% confidence intervals (CI).
Results A total of 20858 patients were included for causal discovery, and the estimation was performed on 8,130 complete-case observations. The set of direct causes of severe fibrosis (excluding diabetes) identified by the causal discovery algorithm comprised 9 variables: diabetes, geographical origin, age, type of hepatitis, total cholesterol, HDL cholesterol, past alcohol consumption, blood glucose, and sex, which constituted the adjustment set. The CDE of diabetes on severe fibrosis among patients with viral hepatitis was significantly positive, with an estimated odds ratio of 1.63 (95% CI [1.40, 1.90]).
Conclusions Even after adjusting for confounding using a targeted, data-driven approach, diabetes still appears to have a direct and statistically significant effect on liver fibrosis in patients with chronic viral hepatitis. This suggests that managing diabetes may be a direct lever for mitigating fibrosis progression in patients with chronic viral hepatitis, not just an indirect one through other metabolic and lifestyle factors.

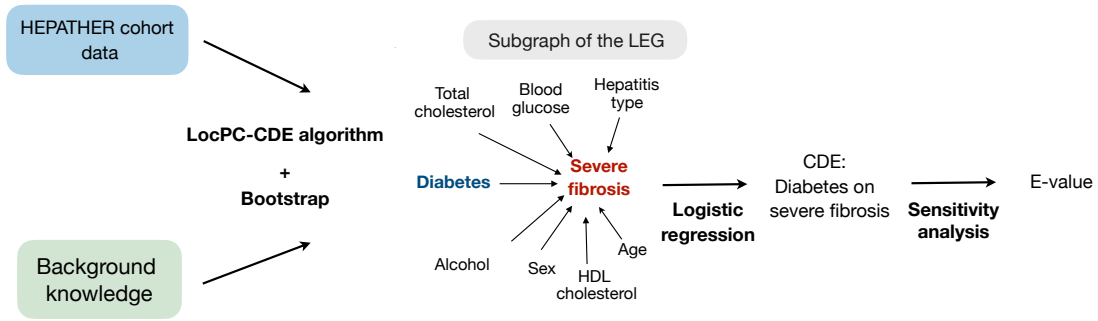


Figure 1: Schematic representation of the procedure for estimating the CDE of diabetes in severe fibrosis using LocalPC-CDE algorithm.

1 INTRODUCTION

Randomized trials are often infeasible or unethical in epidemiology, making observational data the primary resource for causal investigation. To draw valid conclusions from such data, causal inference frameworks such as structural causal models (SCMs) and directed acyclic graphs (DAGs) provide formal tools for estimating causal effects under explicit assumptions [25, 37, 26]. Of particular interest is the *controlled direct effect* (CDE), which captures the effect of an exposure on an outcome that is not transmitted through mediating variables.

When the underlying DAG is unknown, *causal discovery* methods can help infer it directly from the data [32, 30, 13, 2, 3]. Unlike classical approaches that aim to recover the entire DAG, *local* causal discovery restricts attention to the part of the graph relevant for a given question. This reduces assumptions, improves computational efficiency, and yields valid adjustment sets for causal effect estimation, including the CDE.

In this study, we apply the local causal discovery algorithm LocalPC-CDE [21] to baseline data from the French *ANRS CO22 HEPATHER* cohort, which collects detailed clinical, biological, and virological information on patients with chronic viral hepatitis. Chronic infection with hepatitis B (HBV) or hepatitis C (HCV) is a major cause of liver-related morbidity, due to the risk of progression to cirrhosis, hepatocellular carcinoma, and liver failure [19, 20, 8, 10]. Liver fibrosis, the accumulation of extracellular matrix proteins in response to chronic injury, is a key determinant of prognosis and is staged from F0 (no fibrosis) to F4 (cirrhosis), with severe fibrosis defined as stage F3–F4.

Diabetes mellitus, a chronic metabolic disorder characterized by hyperglycemia due to impaired insulin secretion or action, is common among patients with viral hepatitis. Mechanistic evidence suggests that diabetes may accelerate fibrosis progression through insulin resistance, hepatic fat accumulation, oxidative stress, protein glycation, and mitochondrial dysfunction [12, 23, 42]. Observational studies have reported associations between diabetes and advanced fibrosis [19, 20, 8, 10], but whether diabetes exerts a *direct causal effect* on liver fibrosis—independently of other mediating factors—remains unclear. Using local causal discovery, combined with robustness checks such as bootstrap augmentation [11] and sensitivity analyses [39], we estimate the CDE of diabetes on severe fibrosis, showing that indeed diabetes exerts directly affects liver fibrosis.

Beyond this specific application to viral hepatitis, the broader aim of this paper is to illustrate how local causal discovery can complement, rather than replace, traditional epidemiological approaches for causal analysis. By focusing only on the parts of the DAG relevant to a given question, local discovery provides a practical way to identify valid adjustment sets and quantify direct effects, especially when the full DAG is uncertain. In doing so, it offers epidemiologists a flexible tool for strengthening causal claims in observational research, particularly in settings where randomized trials are not feasible.

To make our approach more transparent, Figure 1 presents a pipeline that visually summarizes the steps we followed to estimate the controlled direct effect of diabetes on severe fibrosis in the HEPATHER cohort.

$$X \longleftarrow Z \longrightarrow Y \qquad X \longrightarrow Z \longrightarrow Y \qquad X \longrightarrow Z \longleftarrow Y$$

(a) Confounder. (b) Mediator. (c) Collider.

Figure 2: Illustration of three fundamental causal structures. (a) A confounder, where a common cause Z of both exposure X and outcome Y induces spurious associations. To estimate the CDE of X on Y, it is essential to adjust for such confounders. (b) A mediator, where part of the effect of X on Y is transmitted through an intermediate variable Z. For the CDE, one must adjust for the mediator to isolate the direct pathway. (c) A collider, where two variables X and Y both cause a third variable Z. Adjusting on Z induces a spurious association between X and Y (selection bias) and should therefore be avoided.

2 MATERIEL AND METHODS

2.1 Controlled Direct Effect

Let $\mathbf{Z}$ denote the set of the p observed direct causes of fibrosis (except diabete) in the estimated local causal structure, $D \in \{0,1\}$ the binary variable indicating diabetes status (1 = diabetic, 0 = healthy), and $F \in \{0,1\}$ the binary variable indicating severe fibrosis (1 = stage $\geq$ F3, 0 = stage $<$ F3).

We aim to estimate the *direct effect* of having diabetes on the risk of severe liver fibrosis among patients with viral hepatitis. This effect is considered *direct* because we want to isolate it from any potential mediating pathways. A *mediator* is a variable that lies on the causal path from diabetes to fibrosis (e.g., a variable M such that $D \to M \to F$). If such mediators exist, our goal is to estimate the effect of diabetes on fibrosis that does *not* operate through them—this is referred to as the *direct effect* along the path $D \to F$. One way to formalize and estimate this quantity is through the *Controlled Direct Effect* (CDE). The CDE represents the effect of setting diabetes to a fixed value while holding all other direct causes of fibrosis at specific values. From a causal perspective, this corresponds conceptually to an intervention where diabetes is set to $D = 1$ (denoted $do(D = 1)$), and all other direct causes of fibrosis (denoted $\mathbf{Z}$) are fixed to specific values $\mathbf{Z} = \mathbf{z}$ via $do(\mathbf{Z} = \mathbf{z})$. Then, we compare this to a scenario where diabetes is set to $D = 0$ while keeping the same intervention on $\mathbf{Z}$. The CDE of diabetes (D) on severe fibrosis (F) can then be quantified using an odds ratio comparing the two interventional distributions:

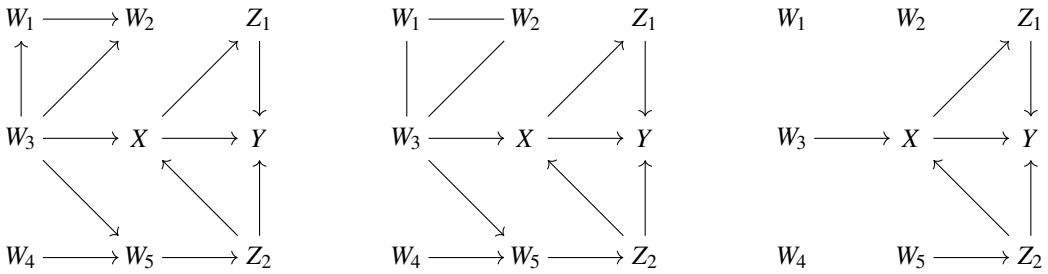
$$OR(D \to F) = \frac{P(F = 1 \mid do(D = 1), do(\mathbf{Z} = \mathbf{z}))/P(F = 0 \mid do(D = 1), do(\mathbf{Z} = \mathbf{z}))}{P(F = 1 \mid do(D = 0), do(\mathbf{Z} = \mathbf{z}))/P(F = 0 \mid do(D = 0), do(\mathbf{Z} = \mathbf{z}))}.$$

We now turn to the question of how to estimate this causal effect from observational data. Three patterns must be considered: (1) confounders (variables causing both diabetes and severe fibrosis), which may bias the estimation if not adjusted for; (2) mediators (intermediate variables caused by diabetes and in turn causing fibrosis), which must be adjusted for to estimate the direct effect; and (3) colliders [25] (variables caused by both diabetes and fibrosis), which should not be adjusted for, as doing so would introduce selection bias [17, 35]. Figure 2, presents a visual illustration of confounders, mediators and colliders. Conditioning on all direct causes of fibrosis simultaneously blocks bias due to confounders and influence of mediators, while preventing adjustment for colliders (which, by definition, are consequences of fibrosis). Under the assumption of no unmeasured confounding, the set of all direct causes of fibrosis thus constitutes a causally valid adjustment set, in the sense that the causal odds ratio $OR(D \to F)$ can be identified from observational data. Denoting by $\mathbf{Z}$ the set of direct causes, identifiability is expressed by the equivalence between interventional and conditional probabilities[1]:

$$P(F = 1 \mid do(D = d), do(\mathbf{Z} = \mathbf{z})) \;=\; P(F = 1 \mid D = d, \mathbf{Z} = \mathbf{z}).$$

Remark that the set $\mathbf{Z}$ is not only valid but also optimal in the sense that it minimizes the estimation variance [16]. To estimate the CDE by adjusting on $\mathbf{Z}$ from finite data we use a logistic regression model, where $\widehat{OR}(D \to F) = \exp(\hat{\beta}_{D \to F})$. The reported 95% confidence intervals are computed as $[\exp(\hat{\beta} \pm 1.96 \cdot \widehat{\sigma}_{\beta})]$, where $\widehat{\sigma}_{\beta}$ denotes the estimated standard error of the coefficient $\widehat{\beta}$. The models are fitted on the complete cases using the variables identified through causal discovery within the adjustment set.

[1] It is important to note that, in general, one cannot simply remove the *do* operator from an interventional expression to obtain a valid conditional expression. This equivalence only holds when conditioning on the specific set $\mathbf{Z}$, which contains all direct causes of severe fibrosis and therefore blocks (under the assumption of no unmeasured confounding) all causal paths from diabetes to fibrosis.



(a) True DAG. (b) Output of global causal discovery. (c) Output of local causal discovery.

Figure 3: Illustration of the difference between global and local causal discovery. (a) The underlying true DAG. (b) The output of PC, a global causal discovery algorithm, which attempts to recover the entire graph and in this example requires around 490 conditional independence tests. (c) The output of LocPC-CDE, a local causal discovery algorithm focused only on the part of the graph relevant for estimating a CDE, requiring only about 290 tests.

To assess the sensitivity of $\widehat{OR}(D \to F)$ to potential unmeasured confounding between diabetes and severe fibrosis, we compute the associated *E-value* [39]. The E-value quantifies the minimum strength of association that an unmeasured confounder would need to have with both the exposure and the outcome, conditional on the measured covariates, in order to fully explain away this estimated effect.

The main difficulty thus lies in identifying such a set of variables $\mathbf{Z}$. The use of Directed Acyclic Graphs (DAGs) is a powerful tool for this purpose: the required variables are simply those with a directed edge pointing to F in the DAG. If the DAG is known in advance (e.g., based on expert knowledge), then this identification is straightforward. However, the DAG may also be partially or completely unknown. In such cases, (local) causal discovery becomes a useful approach.

2.2 Local Causal Discovery Procedure

If the underlying DAG were known in advance, it would be possible to identify the set of direct causes of the target variable, severe fibrosis, by simply selecting all its parent nodes in the graph. This would allow direct estimation of the CDE by adjusting for these variables. However, we assume here that the DAG is not fully known a priori and must be inferred from the data. This task—known as *global causal discovery*—is both computationally and theoretically challenging. Recent approaches suggest that instead of recovering the entire DAG, it is often sufficient to estimate only a *local* portion of the graph surrounding the target (called a Local Essential Graph, LEG). Such local structures are enough to identify the direct causes of the target variable, and they can be learned more efficiently and robustly. This targeted approach is referred to as *local causal discovery*, and it forms the basis of the method we use to identify valid adjustment sets for CDE estimation. Figure 3 illustrates the difference between global and local causal discovery.

Local Causal Discovery Algorithm Local causal discovery is performed with severe fibrosis as the target variable by using the LocPC-CDE algorithm [21] for causal discovery. We briefly describe here the functioning of the LocPC-CDE algorithm; for a detailed explanation and theoretical guarantees, see [21]. This method relies on conditional independence testing, following principles similar to the well-known Peter-Clark (PC) algorithm [32] introduced for global causal discovery (see [31, 27, 7, 4] for applications of the global PC algorithm to epidemiological data). LocPC-CDE begins by testing for conditional independencies between the target variable (here, severe fibrosis) and all other variables to identify its direct neighbors (i.e., direct causes and effects). Then, edge orientations are inferred using conditional independence information combined with logical orientation rules (Meek's rules [22]). The goal is to orient all edges adjacent to the target. The more of the graph is discovered, the more orientations become possible. To facilitate this, local discovery is iteratively extended to nodes adjacent to the target, then to their neighbors, and so on—progressively revealing and orienting the *neighborhood* of the target. The algorithm stops automatically (and optimally, in the sense that it minimizes the neighborhood size and thus the number of tests and potential errors) once all edges adjacent to the target are oriented[2]. The algorithm is, in principle,

[2]It is possible that this condition is never met, in which case the algorithm returns that the direct effect is not identifiable.

fully nonparametric, but it requires an appropriate conditional independence test. To ensure reasonable computation time and a consistent testing procedure, we choose to discretize continuous variables in order to apply a parametric independence test. Discretization is performed using clinical thresholds as detailed in Table 1. Conditional independence is then assessed using the G^2 test (likelihood ratio test for categorical variables [33]). We use the exact same Python implementation of the LocPC-CDE algorithm as in [21]. The LocPC-CDE algorithm is consistent in recovering the correct LEG under the assumptions of local causal sufficiency (absence of hidden confounding factors involving any variable on which discovery is performed) and local faithfulness (a technical, untestable assumption shared by all causal discovery algorithms based on conditional independence tests, here relaxed to apply only locally).

Significance Threshold As previously discussed, the LocPC-CDE algorithm relies on conditional independence tests to infer the LEG in the vicinity of the target variable. This requires specifying a significance threshold α, which determines the rejection criterion for the null hypothesis of conditional independence. Note that, a significance threshold α does not represent a significance level for the estimated graph, as the LocPC-CDE algorithm conduct a sequence of conditional independence tests, where the outcome of each test depends on the results of the previous tests. However, α can be considered as a sparsity level. A larger value of α tends to produce a denser estimated local graph, with more edges retained, while a smaller α yields a sparser graph, with fewer edges included. In our analysis, we adopt the conventional threshold of $\alpha = 0.05$.

Background knowledge Causal discovery algorithms can generally incorporate background knowledge, that is, prior information on certain causal relationships (e.g., derived from expert opinion, the literature, or previous studies). Such knowledge can take different forms, but in this work we focus on orientation constraints: we pre-specify that some edges, if present, must be oriented in a certain direction. Importantly, this does not force the algorithm to include the edge, but rather ensures that if it is detected, its orientation is consistent with prior knowledge. Incorporating background knowledge in this way can reduce computational time and improve algorithmic performance, provided that the prior information is correct. Moreover, these constraints do not need to concern only liver fibrosis: any reliable orientation can improve the learning process. In our application, we impose some of such constraints. First, no variable in the model can causally influence age, sex, or geographical origin; therefore, if an edge connects to these variables, it cannot be oriented towards them (e.g., liver fibrosis cannot cause age). Second, diabetes is constrained to be a cause of severe fibrosis, given consistent evidence from both epidemiological data and biological mechanisms. Third, by definition, cirrhosis is a consequence of severe fibrosis, and hepatocellular carcinoma is a consequence of fibrosis (and potentially cirrhosis); thus, any edges among these variables must point from fibrosis towards cirrhosis or carcinoma. Finally, for viral hepatitis, we specify that the type of infection is viral and then is not caused by the measured variables of the model, except for potential causes from geographical origin (e.g., higher prevalence in specific regions[41]), sex (e.g., behavioral risk factors[14]), and age[41].

Missing data The local causal discovery algorithms we use assume complete data for correct operation, and missing values can potentially introduce bias, especially if the missingness is related to other variables. To handle this, we assume that the data are *Missing Completely At Random* (MCAR) and apply a *test-wise deletion* strategy [34]. That is, for each statistical test, we use only individuals with observed values for the variables involved in that test, while keeping individuals with missing values in unrelated variables. For instance, when testing the conditional independence of X and Y given $\mathbf{Z}$, we exclude rows with missing values in X, Y, or any variable in $\mathbf{Z}$, but retain rows with missing values in other variables. Under the MCAR assumption, this approach preserves the validity of the discovery algorithms and maximizes the statistical power of each test.

Uncertainty on the inferred causal structure If we run the LocPC-CDE algorithm as is, we obtain a point estimate of the LEG, which is subject to finite-sample bias. Such bias may affect conditional independence tests, potentially leading to spurious edges, missing true edges, or incorrect edge orientations. The main issue is the absence of an uncertainty measure: would a slightly different sample have led to the same result? To what extent is a given edge driven by our specific sample? To address these questions,

a common suggestion is to incorporate a bootstrap approach into the discovery process (an approach suggested in [11]). For a sample size N, we draw B bootstrap samples of size N, run the local causal discovery algorithm on each sample, and measure the frequency of a feature of interest across the B outputs. We apply this procedure exclusively in our real-data application. In this context, when estimating the CDE of diabetes on severe fibrosis, our interest lies in identifying the set of parents (direct causes) of severe fibrosis (denoted F). We therefore assess, for a given N, the confidence that Z is a direct cause of F as:

$$p_N(F) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{Z \in Pa(F, \widehat{\mathscr{L}}_b^N)}, \tag{1}$$

where $\mathbb{1}$ is the indicator function and $\widehat{\mathscr{L}}_b^N$ is the b-th inferred LEG from sample of size N. Since the algorithm is consistent for recovering the LEG, this quantity should converge to 1 for true direct causes of fibrosis, and to 0 for non-causes as N increases. We use the full dataset size in the bootstrap procedure $N = 20858$, and set $B = 30$ due to the computational cost of the causal discovery method. Variables are selected as direct causes of severe fibrosis if they are identified in more than 50

2.3 Data

Data were obtained from the ANRS CO22 HEPATHER[3] cohort, which is a French national, multicenter, prospective, observational cohort study of patients with hepatitis B infection or hepatitis C (past or present) infection [28]. Initiated in 2012, this prospective cohort follows participants for approximately ten years to improve understanding and management of new hepatitis treatments and currently includes over 20,000 patients, see details in [28]. In this study, we used sociodemographic, clinical, and biological data collected at the time of cohort enrollment from patients with chronic or resolved hepatitis B or C.

Variables We selected 21 baseline variables potentially related to liver fibrosis, excluding direct fibrosis biomarkers (e.g., gammagt, albumin, alat, etc.) to avoid deterministic or near-deterministic relations that could bias causal discovery, as they would violate the local faithfulness assumption and might prevent detecting the true causal relationships[4]. These variables were: (1) the outcome *severe fibrosis* (fibrosis stage F3 or higher); (2) the exposure *diabetes* (diagnosis of diabetes); (3) other variables *age* (years at inclusion), *sex* (biological sex), *current smoker* and *ever smoked* (tobacco use status), *current* and *history of excessive alcohol consumption*, *body mass index* (BMI, weight-to-height ratio), *HBP* (history of high blood pressure), *cirrhosis* (most advanced stage of liver fibrosis), *hemoglobin* (blood oxygen-carrying capacity), *neutrophils* (immune cell count), *creatinine* (renal function marker), *blood glucose* (glycemia), *total cholesterol* (overall blood cholesterol concentration), *HDL cholesterol* ("good" cholesterol fraction), *triglycerides* (blood lipid level), *hepatocellular carcinoma at inclusion* (HCC, liver cancer status), *hepatitis type* (B,C or BC), and *geographical origin* (region of birth). Since the dataset is mixed (binary and continuous variables), all variables were discretized using clinically relevant thresholds [18] to allow a G^2 conditional independence test (Table 1). All baseline data were retained ($N = 20858$), including cases with missing values, as many conditional independence tests do not involve the missing variables, thus preserving statistical power.

Models fitting The local causal discovery algorithm with bootstrap is then applied to this sample to identify the likely direct causes of severe fibrosis. If the CDE is identifiable, the logistic regression model is fitted on the complete cases for the outcome and selected direct causes, yielding the causal odds ratio estimate. Our model, adjusted on a causally valid set, is compared to two non-causal statistical models. The first is a naive model, adjusting for no covariates and assessing the univariate association between diabetes and severe fibrosis. The second is an overadjusted model, controlling for all variables in the dataset. Each model is fitted on the complete data corresponding to its included variables. All models are logistic regressions. Furthermore, although local causal discovery is performed on discretized data, the original continuous measurements are available. To strengthen the robustness of our findings, we will also fit a logistic regression model including the continuous variables alongside the model derived from the discretized dataset.

[3] `https://anr.fr/ProjetIA-11-EQPX-0021`

[4] Intuition: if a variable (the proxy) is a quasi-deterministic function of fibrosis, it renders nearly all other variables conditionally independent of it, since almost all information about fibrosis is contained in this proxy.

Variable	Threshold	n (1)	% (1)	n (0)	% (0)	NA
Diabetes	1 = yes	2609	12.5	18238	87.5	11
Severe fibrosis (F3 or higher)	F3, F3/F4, F4	7248	40.2	10794	59.8	2816
Age	$\geq$ 50 years	7409	35.5	13446	64.5	3
Sex	1 = male	12367	59.3	8490	40.7	1
Current smoker	1 = yes	6207	29.8	14601	70.2	50
Ever smoked	1 = yes	11503	55.3	9305	44.7	50
Current excessive alcohol consumption	1 = yes	175	1.1	15809	98.9	4874
History of excessive alcohol consumption	1 = yes	4950	23.7	15893	76.3	15
Body mass index (BMI)	$\geq$ 29.9 kg/m^2	17726	85.8	2939	14.2	193
HBP	1 = yes	5646	27.1	15180	72.9	32
Cirrhosis	1 = yes	5279	25.4	15515	74.6	64
Hemoglobin	$\geq$ 12 g/dL	18056	91.4	1692	8.6	1110
Neutrophils	$\geq$ 1500 /mm^3	17653	93.1	1304	6.9	1901
Creatinine	$\geq$ 13 mg/L	17586	94.6	1012	5.4	2260
Blood glucose	$\geq$ 1.1 g/L	3452	21.3	12758	78.7	4648
Total cholesterol	$\geq$ 2.5 g/L	13186	93.4	926	6.6	6746
HDL cholesterol	$\geq$ 0.5 g/L	5391	52.4	4900	47.6	10567
Triglycerides	$\geq$ 1.5 g/L	11525	82.8	2396	17.2	6937
Hepatocellular carcinoma (at inclusion)	1 = yes	1164	5.6	19694	94.4	0
Hepatitis Type:						
B		6192	29.7			
B-C (coinfection)		2312	11.1			
C		12354	59.2			
Geographical origin:						
France		10966	52.6			
Sub-Saharan Africa		2993	14.3			
North Africa		2078	10.0			
Asia		1364	6.5			
French Overseas Territories		546	2.6			
Eastern Europe		871	4.2			
Western Europe		259	1.2			
Southern Europe		1320	6.3			
Other		212	1.0			
Missing		249	1.2			

Table 1: Variables on which local causal discovery is performed ($N = 20858$).

For the implementation, we used a Python (v3.13.3) implementation of LocPC-CDE, and R (v4.5.0) for fitting logistic regression models (`glm`) and for computing the E-value (package `EValue`).

3 RESULTS

Recall that answering the question *What is the CDE of diabetes on severe fibrosis in patients with viral hepatitis?* requires: (i) discovering the direct causes of severe fibrosis in our dataset, and (ii) estimating the causal odds ratio of diabetes by adjusting for these other direct causes.

(i) Estimation of direct causes The bootstrap procedure output is shown in Figure 4, which displays the proportion of bootstrap samples in which each variable is identified as a direct cause of severe fibrosis. Some variables show strong consensus as direct causes: diabetes, age, geographic origin, hepatitis type, and sex are identified as such in over 90% of the samples. Blood glucose, HDL cholesterol, and alcohol consumption history follow. We retain all these variables as direct causes, as they are identified as such in more than 50% of the samples.

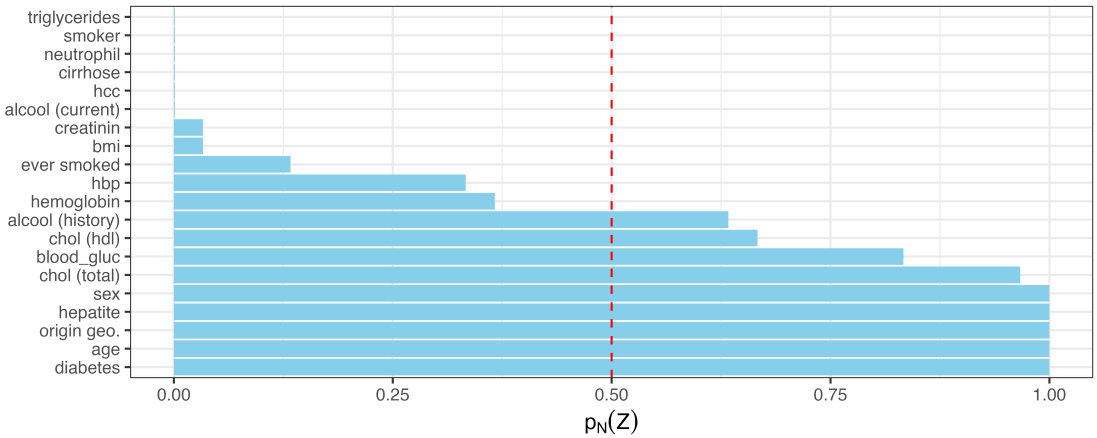


Figure 4: Confidence $p_N(Z)$ that each variable Z is a direct cause of severe fibrosis. Red = 50%. $\alpha = 0.05$, $B = 30$.

Other variables are rarely or never identified as direct causes of severe fibrosis: triglycerides and smoking are never direct causes, BMI and creatinine only rarely, while hemoglobin and hypertension are occasionally flagged. We attribute these occasional assignments to finite-sample noise and the algorithm's sensitivity to minor errors that can induce incorrect edge orientations. Notably, our method only constrains edge orientations based on background knowledge, not the presence of edges themselves. For example, diabetes is identified as a direct cause of fibrosis in 100% of bootstrap samples, indicating that no sample showed a conditional independence between diabetes and severe fibrosis, confirming a robust causal link. These findings for the other odds ratios are consistent with the existing literature: age, past alcohol consumption, and sex have been associated with severe fibrosis in patients with viral hepatitis [29]; the link between severe fibrosis and glucose levels, including independently of diabetes, has been identified as a risk factor for liver fibrosis [6]; and low HDL cholesterol is a known risk factor for liver disease [9].

Thus, the set of variables $\mathbf{Z} =$ {age, geographic origin, hepatitis type, sex, total and HDL cholesterol, alcohol consumption history, blood glucose} forms our presumed causally valid adjustment set for the logistic model.

(ii) CDE estimation The model adjusted on the adjustment set comprising the direct causes of severe fibrosis is presented in Table 2. The causal odds ratio corresponding to the CDE of diabetes on severe fibrosis is **1.63 (95% CI [1.40; 1.90])**, indicating that, among patients with hepatitis B or C, having diabetes *causes* a 1.63-fold increase (95% CI [1.40; 1.90]) in the odds of severe fibrosis.

While this addresses the primary question of this study, it is noteworthy that the adjustment set would be the same to estimate the CDE of any other variable on severe fibrosis, as the direct causes of fibrosis constitute a set that is independent of the exposure under study. Consequently, all odds ratios in Table 2 can be interpreted causally as CDE odds ratios. In this particular case (since we are studying direct effects), the table avoids the Table 2 Fallacy [40], which typically cautions that coefficients other than the exposure of interest in a multivariable regression cannot be interpreted causally. These findings regarding the other odds ratios are consistent with the existing literature that has previously highlighted these associations: for age [29], and for blood glucose (notably independent of diabetes, for example at the prediabetes stage) [24, 6].

The sensitivity analysis using the E-value [39] yielded an estimate of 3.50 (lower CI bound of 2.72) expressed in odds ratios[5]. This implies that, to reduce the observed odds ratio for the effect of diabetes on severe fibrosis to 1, an unmeasured confounder would need to be associated with both diabetes and severe fibrosis with an odds ratio of at least 3.50 each (2.72 for the lower bound of the confidence interval). The E-value does not define a robustness threshold; its interpretation relies on expert judgment regarding the plausibility of such unmeasured confounding. In our context, such an effect size—ranging from 2.72 to 3.50—would be among the largest odds ratios observed in our logistic regression models, comparable in

[5]The `EValue` package computes the E-value from an odds ratio by first approximating it with a risk ratio using a square-root transformation [38]. We applied the reverse transformation (squaring the reported value) to express the E-value in odds ratios for consistency with our results.

Variable	OR [95% CI]
diabetes	1.63* [1.40; 1.90]
(Intercept)	0.11* [0.09; 0.13]
sex (male)	1.53* [1.37; 1.71]
age	2.30* [2.04; 2.59]
geographical_origin: AFNOR	1.34* [1.07; 1.69]
geographical_origin: ASIE	1.32 [1.00; 1.74]
geographical_origin: OTHER	1.13 [0.66; 1.89]
geographical_origin: DOMTOM	1.50* [1.11; 2.04]
geographical_origin: EUREST	1.43* [1.04; 1.96]
geographical_origin: EUROUEST	1.15 [0.72; 1.86]
geographical_origin: EURSUD	1.51* [1.17; 1.95]
geographical_origin: FRA	1.23* [1.02; 1.48]
hepatite: BC	3.43* [2.85; 4.12]
hepatite: C	3.00* [2.59; 3.49]
excessive alcohol consumption history	1.54* [1.38; 1.73]
blood glucose	1.48* [1.30; 1.68]
chol total	0.49* [0.39; 0.60]
chol hdl	0.57* [0.51; 0.63]
AIC = 9670, BIC = 9796, Log Lik. = −4817, Dev. = 9634, $N = 8132$	

* 0 outside the confidence interval.

Table 2: Logistic Regression with Odds Ratios.

magnitude to the effect of hepatitis type, suggesting that only a strong unmeasured confounder could fully explain away the positive association.

Finally, Figure 5 compares our estimation with two other models (no adjustment and adjustment on all variables) as well as a model using continuous rather than discretized data. First, we observe that the estimation is insensitive to the variable type in our case: the odds ratios with and without discretization are not significantly different, and their confidence intervals largely overlap. The naive simple regression model substantially overestimates the odds ratio of diabetes on fibrosis, further confirming the need for adjustment. The over-adjusted model tends to underestimate the direct effect of diabetes, rendering it nearly non-significant (lower bound of the confidence interval at 1). This difference in point estimates can be explained by adjustment on potential colliders. The wider confidence interval compared to our model can be attributed to two factors: (i) fewer data points are available in this estimation because it includes more variables, resulting in more missing data, and (ii) the causal adjustment set used in our model is optimal in the sense that it is proven to minimize estimation variance. For all these reasons, our estimation appears to be the most precise and reliable from a causal perspective.

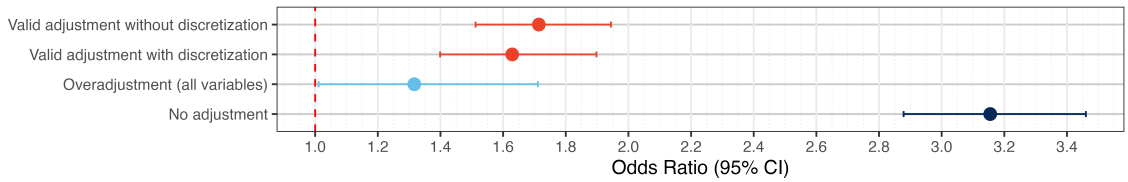


Figure 5: Estimated odds ratios for diabetes on severe fibrosis using the causal adjustment set, comparing discretized vs. original data, a naive unadjusted model, and an over-adjusted model.

4 DISCUSSION

In this study, we demonstrated the use of local causal discovery to estimate the CDE of diabetes on severe liver fibrosis in patients with chronic viral hepatitis. Traditional causal inference approaches often rely on

expert-specified DAGs, which may be incomplete or partially unknown in complex epidemiological settings. Local causal discovery, by focusing only on the relevant portion of the graph necessary for a specific causal question, allows for data-driven identification of valid adjustment sets, reducing the assumptions and computational burden typically required by global causal discovery methods.

Our application to the *ANRS CO22 HEPATHER* cohort to estimate the direct effect of diabetes on severe liver fibrosis among patients with chronic hepatitis B, C, or B–C co-infection, showed that local discovery causal be useful to identify a correct adjustment set. Among 20 candidate variables, the algorithm identified 9 adjustment variables. Logistic regression yielded a controlled direct effect of diabetes on severe fibrosis with an odds ratio of 1.51 (95% CI: 1.40–1.90), suggesting a moderate but statistically significant effect. This application confirms the practical feasibility of local causal discovery in a real-world epidemiological dataset with multiple covariates and potential mediators, especially when it is used with robustness checks, such as bootstrapping and sensitivity analyses, are used to assess the stability of the estimated effects and the influence of potential unmeasured confounding.

Nevertheless, some limitations warrant consideration. Interpretation of the results presented in this paper relies on several strong assumptions. First, **local causal sufficiency and local faithfulness** are assumed, meaning that all relevant variables are included locally and observed dependencies reflect true causal relationships. Violations of these assumptions, plausible in biological systems (about the strong assumptions underlying causal discovery methods and the consequences of their violation, see [36, 1, 3, 5]), may lead to missing or unoriented edges and affect the identifiability of the causal effect. Second, data are assumed to be **missing completely at random (MCAR)**; if missingness depends on measured factors, spurious associations may bias the estimated effect. It is important to note that all true edges of the local graph are recovered in the presence of missing data even if the MCAR assumption is violated, although additional spurious edges may also be discovered. The MCAR assumption can be relaxed by performing global causal discovery [34], but this comes at a computational cost (in particular for a bootstrap-based approach as considered here), and global causal discovery generally performs worse than local causal discovery for the task of direct effect identification [21]. Moreover, as in standard causal inference, we assume an underlying **acyclic causal structure**; feedback loops or cycles in biological processes could limit the validity of the conclusions.

Another important limitation of this paper is that we use a local causal discovery approach to select the relevant set of parents and then proceed to estimate causal effects using the same data. This practice compromises the coverage guarantees for the confidence intervals of the estimated causal effect [15]. In our study, since we used a test deletion strategy, the data used for causal discovery could differs with the data used for the causal effect estimation. While this approach is likely to mitigate the issue, we cannot formally guarantee that it fully resolves it.

In conclusion, this study highlights that local causal discovery can serve as a valuable tool for causal inference in epidemiology, particularly when its limitations are carefully considered and appropriate robustness checks are applied. We emphasize that these methods are not intended to replace traditional approaches or expert knowledge, but rather to complement them, offering actionable insights for public health and clinical decision-making in situations where conventional methods alone may be insufficient.

Acknowledgements

The ANRS CO22 Hepather cohort received funding from Inserm-ANRS, Agence Nationale de la Recherche (ANR-11-EQPX-0021, ANR-19-COHO-0002), Direction Générale de la Santé (DGS), and Gilead, MSD, Abbvie, Janssen, BMS, Roche. This work was also supported by the CIPHOD project (ANR-23-CPJ1-0212-01). We would like to thank the study participants and the participating clinicians at each site.

References

[1] Holly Andersen. When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, (5):672–683, 2013.

[2] Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Entropy-based discovery of summary causal graphs in time series. *Entropy*, 24(8), 2022.

[3] Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *J. Artif. Int. Res.*, 73, apr 2022.

[4] V. Asvatourian, C. Coutzac, N. Chaput, C. Robert, S. Michiels, and E. Lanoy. Estimating causal effects of time-dependent exposures on a binary endpoint in a high-dimensional setting. *BMC Medical Research Methodology*, 18(1):67, July 2018.

[5] Ali Aït-Bachir, Charles K. Assaad, Christophe de Bignicourt, Emilie Devijver, Simon Ferreira, Eric Gaussier, Hosein Mohanna, and Lei Zan. Case studies of causal discovery from it monitoring time series, 2023. The History and Development of Search Methods for Causal Structure Workshop at the 39th Conference on Uncertainty in Artificial Intelligence.

[6] Xinxia Chang, Hua Bian, Mingfeng Xia, Xiaopeng Zhu, Xiaoyang Sun, Xinyu Yang, Jian Gao, Huandong Lin, Hongmei Yan, and Xin Gao. Postprandial glucose is correlated with an increasing risk of liver fibrosis in chinese patients with nonalcoholic fatty liver disease. *Diabetes & Metabolism*, 48(6):101377, 2022.

[7] Chelsea Cheek, Huong Zheng, Brian R. Hallstrom, and R. E. Hughes. Application of a causal discovery algorithm to the analysis of arthroplasty registry data. *Biomedical Engineering and Computational Biology*, 9:1179597218756896, Feb 2018.

[8] Sandeep Chhabra, Sukhraj P. Singh, Arshdeep Singh, Varun Mehta, Amninder Kaur, Namita Bansal, and Ajit Sood. Diabetes mellitus increases the risk of significant hepatic fibrosis in patients with non-alcoholic fatty liver disease. *Journal of Clinical and Experimental Hepatology*, 12(2):409–416, 2022.

[9] L. Crudele, C. De Matteis, E. Piccinin, R. M. Gadaleta, M. Cariello, E. Di Buduo, G. Piazzolla, P. Suppressa, E. Berardi, C. Sabbà, and A. Moschetta. Low hdl-cholesterol levels predict hepatocellular carcinoma development in individuals with liver fibrosis. *JHEP Reports*, 5(1):100627, November 2022.

[10] Joana D'Arc Matos França de Abreu, Rossana Sousa Azulay, Vandilson Rodrigues, Sterffeson Lamare Lucena de Abreu, Maria da Glória Tavares, Flávia Coelho Mohana Pinheiro, Clariano Pires de Oliveira Neto, Caio Andrade, Alexandre Facundo, Adriana Guimarães Sá, Patrícia Ribeiro Azevedo, Ana Gregória Pereira de Almeida, Debora Camelo de Abreu Costa, Rogério Soares Castro, Marcelo Magalhães, Gilvan Cortês Nascimento, Manuel dos Santos Faria, and Adalgisa de Souza Paiva Ferreira. Predictors of hepatic fibrosis in type 2 diabetes patients with metabolic-dysfunction-associated steatotic liver disease. *Biomedicines*, 12(11), 2024.

[11] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach, 2013.

[12] Hideki Fujii, Norifumi Kawada, and Japan Study Group Of Nafld Jsg-Nafld. The role of insulin resistance and diabetes in nonalcoholic fatty liver disease. *International Journal of Molecular Sciences*, 21(11):3863, May 2020.

[13] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, Volume 10 - 2019, 2019.

[14] B. Gnyawali, A. Pusateri, A. Nickerson, S. Jalil, and K. Mumtaz. Epidemiologic and socioeconomic factors impacting hepatitis b virus and related hepatocellular carcinoma. *World Journal of Gastroenterology*, 28(29):3793–3802, 2022.

[15] Paula Gradu, Tijana Zrnic, Yixin Wang, and Michael I. Jordan. Valid inference after causal discovery. *Journal of the American Statistical Association*, 120(550):1127–1138, 2025.

[16] Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 03 2022.

[17] M. J. Holmberg and L. W. Andersen. Collider bias. *JAMA*, 327(13):1282–1283, 2022.

[18] L. Lam, H. Fontaine, M. Bourliere, C. Lusivika-Nzinga, C. Dorival, D. Thabut, F. Zoulim, F. Habersetzer, T. Asselah, J. C. Duclos-Vallee, J. P. Bronowicki, P. Mathurin, T. Decaens, N. Ganne, D. Guyader, V. Leroy, I. Rosa, V. De Ledinghen, P. Cales, X. Causse, D. Larrey, O. Chazouilleres, M. Gelu-Simeon, V. Loustaud-Ratti, S. Metivier, L. Alric, G. Riachi, J. Gournay, A. Minello, A. Tran, C. Geist, A. Abergel, F. Raffi, L. D'Alteroche, I. Portal, N. Lapidus, S. Pol, F. Carrat, and ANRS/AFEF Hepather study group. Predictive factors for hepatocellular carcinoma in chronic hepatitis b using structural equation modeling: a prospective cohort study. *Clinical Research in Hepatology and Gastroenterology*, 45(5):101713, September 2021. Epub 2021 Apr 27.

[19] Xu Li, Yuan Jiao, Yunlong Xing, and Pujun Gao. Diabetes mellitus and risk of hepatic fibrosis/cirrhosis. *BioMed Research International*, 2019(1):5308308, 2019.

[20] Romina Lomonaco, Eddison Leiva, Fernando Bril, Sulav Shrestha, Lydia Mansour, Jeff Budd, Jessica Romero, Siegfried Schmidt, Ku-Lang Chang, George Samraj, John Malaty, Katherine Huber, Pierre Bedossa, Srilaxmi Kalavalapalli, Jonathan Marte, Diana Barb, Danielle Poulton, Nada Fanous, and Kenneth Cusi. Advanced liver fibrosis is common in patients with type 2 diabetes followed in the outpatient setting: The need for systematic screening. *Diabetes Care*, 44:dc201997, 12 2020.

[21] Timothée Loranchet and Charles K. Assaad. Local markov equivalence and local causal discovery for identifying controlled direct effects. In *UAI 2025 Workshop on Causal Abstractions and Representations*, 2025.

[22] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[23] J. Mohamed, A. H. Nazratun Nafizah, A. H. Zariyantey, and S. B. Budin. Mechanisms of diabetes-induced liver damage: The role of oxidative stress and inflammation. *Sultan Qaboos University Medical Journal*, 16(2):e132–e141, May 2016.

[24] J. Park, H. J. Kwon, W. Sohn, J. Y. Cho, S. J. Park, Y. Chang, S. Ryu, B. I. Kim, and Y. K. Cho. Risk of liver fibrosis in patients with prediabetes and diabetes mellitus. *PLOS ONE*, 17(6):e0269070, 2022.

[25] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000.

[26] Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[27] Anne H Petersen, Merete Osler, and Claus T Ekstrøm. Data-driven model building for life-course epidemiology. *American Journal of Epidemiology*, 190(9):1898–1907, 03 2021.

[28] Stanislas Pol, Marc Bourliere, Sandy Lucier, Christophe Hezode, Céline Dorival, Dominique Larrey, Jean-Pierre Bronowicki, Victor DE Ledinghen, Fabien Zoulim, Albert Tran, et al. Safety and efficacy of daclatasvir-sofosbuvir in hcv genotype 1-mono-infected patients. *Journal of hepatology*, 66(1):39–47, 2017.

[29] Thierry Poynard, Vlad Ratziu, Frédéric Charlotte, Zachary Goodman, John McHutchison, and Janice Albrecht. Rates and risk factors of liver fibrosis progression in patients with chronic hepatitis c. *Journal of Hepatology*, 34(5):730–739, 2001.

[30] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel Mahecha, Jordi Muñoz, Egbert Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(2553), 2019.

[31] Manas Sinha, Petter Haaland, Akshay Krishnamurthy, et al. Causal analysis for multivariate integrated clinical and environmental exposures data. *BMC Medical Informatics and Decision Making*, 25:27, 2025.

[32] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[33] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 3 2006.

[34] Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1762–1770. PMLR, 16–18 Apr 2019.

[35] T. Tönnies, S. Kahl, and O. Kuss. Collider bias in observational studies. *Dtsch Arztebl Int*, 119(7):107–122, Feb 2022.

[36] Caroline Uhler, Garvesh Raskutti, Peter Buhlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, 41:436–463, 2012.

[37] Tyler J. VanderWeele. Controlled direct and mediated effects: Definition, identification and bounds. *Scandinavian Journal of Statistics*, 38(3):551–563, 2011.

[38] Tyler J. VanderWeele. On a square-root transformation of the odds ratio for a common outcome. *Epidemiology*, 28(6):e58–e60, November 2017.

[39] Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2011.

[40] Daniel Westreich and Sander Greenland. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4):292–298, Feb 2013.

[41] World Health Organization. Hepatitis b. `https://www.who.int/news-room/fact-sheets/detail/hepatitis-b`, 2024. Accessed August 14, 2025.

[42] Sylwia Ziolkowska, Agnieszka Binienda, Maciej Jabłkowski, Jolanta Szemraj, and Piotr Czarny. The interplay between insulin resistance, inflammation, oxidative stress, base excision repair and metabolic syndrome in nonalcoholic fatty liver disease. *International Journal of Molecular Sciences*, 22(20):11128, Oct 2021.