Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220463

OntoBioStat: Supporting Causal Diagram Design and Analysis

Thibaut PRESSAT LAFFOUILHÈRE[a,b,c,1], Julien GROSJEAN[a,d], Jacques BÉNICHOU[b,e], Stefan J. DARMONI[a,d] and Lina F. SOUALMIA[c,d]
[a]*CHU Rouen, Department of Biomedical Informatics, F-76000 Rouen, France*
[b]*CHU Rouen, Department of Biostatistics, F-76000 Rouen, France*
[c]*Normandie Univ, UNIROUEN, LITIS-TIBS EA 4108, F-76000 Rouen, France*
[d]*LIMICS U1142, Sorbonne Université, Paris, France*
[e]*INSERM U1018, CESP, Université Paris-Saclay, Paris, France*

Abstract. Suitable causal inference in biostatistics can be best achieved by knowledge representation thanks to causal diagrams or directed acyclic graphs. However, necessary and sufficient causes are not easily represented. Since existing ontologies do not fill this gap, we designed OntoBioStat in order to enable covariate selection support based on causal relation representations. OntoBioStat automatic ontological causal diagram construction and inferences are detailed in this study. OntoBioStat inferences are allowed by Semantic Web Rule Language rules and axioms. First, statements made by the users include outcome, exposure, covariate, and causal relation specification. Then, reasoning enable automatic construction using generic instances of Meta_Variable and Necessary_Variable classes. Finally, inferred classes highlighted potential bias such as confounder-like. Ontological causal diagram built with OntoBioStat was compared to a standard causal diagram (without OntoBioStat) in a theoretical study. It was found that confounding and bias were not completely identified by the standard causal diagram, and erroneous covariate sets were provided. Further research is needed in order to make OntoBioStat more usable.

Keywords. Causality, Ontology, Statistic, Bias, Variable selection, Decision support techniques

1. Introduction

According to the Medical Subject Heading (MeSH) thesaurus, '[…] Causes are termed **necessary** when they must always precede an effect and **sufficient** when they initiate or produce an effect [...]'.

In causal inference the aim of the statistical analysis is to provide an unbiased causal effect of an exposure of interest on an outcome (e.g., effect of an oral antidiabetic on pancreatic cancer risk), using for example adjustment methods [1]. Causal diagrams are used in order to select the right sets of covariates that should be adjusted for [2]. Causal diagrams (CDs) are qualitative representations of a given study, with variables as nodes and probabilistic causal relations as edges between variables. CD's representation depends on the use case and there is no tutorial or universal rules

[1] Corresponding Author, Thibaut PRESSAT LAFFOUILHÈRE, 37 Boulevard Gambetta, 76000 Rouen, France; E-mail: t.pressat@chu-rouen.fr.

that could help users to build a causal diagram with necessary and sufficient causes in all cases [3,4]. Existing published ontologies such as the Relation Ontology [5] or Radiology Gamuts Ontology [6] do not cover entirely the complexity of the different causal relations needed for covariate selection (i.e. distinction between counterfactual probabilistic, sufficient and necessary causes). We designed the OntoBioStat [7] ontology in order to support covariate selection for causal inference. OntoBioStat was built using expert knowledge corpus, theoretical cases, and literature review in order to address several competency questions. OntoBioStat is a domain ontology that can help users in their tasks of building and understanding causal diagrams.

This paper focuses on two OntoBioStat features: (i) automatic construction of causal diagrams with necessary causes (called in this article ontological causal diagram), and (ii) reasoning on necessary and sufficient causes. It is divided in two parts: first the description of the classes, relations, rules and instances involved in each of the two features, and then a theoretical study relying on necessary and sufficient causes was presented.

2. Materials and Methods

2.1. OntoBioStat

OntoBioStat was built with the Protégé software [8]. The last version is available at https://bioportal.bioontology.org/ontologies/OBS. Reasoning is supported by the Pellet reasoner [9]. OntoBioStat is composed by 53 classes and 33 relations. Here we focused on 24 classes, five object properties, no data properties, 28 instances, and nine rules from OntoBioStat. Indeed, OntoBioStat knowledge representation includes for example interaction and missing data that are not relevant here (inferences are not impacted). The following classes were used: “Exposure_stressor”, “Outcome” and “Covariates”. “Meta_Variable” class groups “Decision-makers” and “Method_of_measurement”. “Theoretical_Variable” class groups “Condition”, “Environment”, “Status”, “Health_Behavior”, “Intervention_Effect”. “Necessary_Variable” class groups “Exist”, “Available”, “Indicated”, “Prescribed”, “Delivered”, “Investigated”, and “Adhered”. Inferred classes presented were “Reverse_Causality” (subClassof “Outcome”), and “Unadjusted_Confounder” (both bias that cannot be corrected using adjustment), “Mediation_Differential_Confounder” and “Confounder-like” (subClassof “Covariate”) (both bias that should be corrected using adjustment). Object properties *Related_to* and his descendants were used to represent unidirectional, bidirectional, and non-directional probabilistic ‘causal’ relations between two instances. Among *Signed* properties, *Contraindication* and *Absolute_Indication* are the two object properties that represent sufficient (deterministic) causal relations. They were named with ‘indication’ term because most of the time the sufficient causes in biomedical research are patients characteristics that contraindicate or impose the use of a particular treatment.

“Necessary_Variable” and “Meta_Variable” were fed with 28 generic instances that could be used for any study. These instances are not involved in any causal relation until first new instances are added and the reasoner activated. Automatic ontological causal diagram (OCD) construction relies on five Semantic Web Rule Language (SWRL) rules. The inferred classes rely on three rules for necessary causes (see example below (1)) and one axiom for the sufficient causes. Several SWRL rules about

causal reasoning were used to infer all *Related_to* descendants based on *isCauseof* statements that are not developed here.

Inverse_Directed_Relation(?x,?y)^Mediator(?y)^Necessary_Variable(?x) -> Mediation_Differential_Confounder (1)

Figure 1. Decision-support pipelines and steps (1: User enters the two first variables names as instances and define their classes: (i) Exposure_stressor and Outcome, (ii) Theoretical_Class, 2: User activates reasoner, 3: Automatic construction using existent generic instances (purple) and causal relations (black), 4: User adds covariates and causal relations (red), 5: User refreshes the reasoner, 6: OntoBioStat provide inferred classes (yellow) and new object properties (blue), 7: Lecture of the results and inferences explanations.)

Decision-support based on OntoBioStat requires several exchanges of information between the biostatistician and the Protégé (Figure 1).

2.2. Theoretical study

The aim was to obtain an unbiased true causal effect between the use of oral antidiabetics (oad 1 versus oad 2) and time to pancreatic neoplasm diagnosis. Covariates included comorbidities, oad side effects, and co treatment. Based on the following statements, an OCD and a CD were built: (i) patient comorbidity contraindicates the use of oad 1 and may cause a pancreatic neoplasm, (ii) co treatment is prescribed with oad 2 and is known to cause pancreatic neoplasm, (iii) side effects more often caused by oad 2 more often lead to medical consultation.

The OCD created and analyzed with OntoBioStat was confronted to a CD. The CD was created without necessary variables provides by OntoBioStat during automatic construction process. The CD reasoning to solve variable selection was based on the back-door criterion algorithm [10] instead of rules and axioms from OntoBioStat. Reverse causality implies a cyclic graph; hence the directed causal relation from Outcome to Exposure was not included in the CD.

3. Results

For more readability, OCD inferences are presented in one truncated diagram excluding some of the necessary variables and inferred object properties (Figure 2). Explanations about the inferences are the following: (i) ‘co treatment’ is a “Confounder-like” variable because ‘co treatment’ *isCauseof* “Outcome” and *hasCause* ‘prescribed_exposure’ that is an “Indirect_Confounder”; (ii) ‘side effects’ is a “Mediation Differential Confounder” because *hasCause* “Exposure” and *isCauseof* “Necessary_Variable” (1); (iii) ‘comorbidity’ is an *Unadjusted Confounder* because *Contraindication* of “Exposure” and *isCauseof* “Outcome”; (iv) “Outcome” is

classified in "Reverse_Causality" because "Outcome" *isIndirectCauseof* "Exposure". The standard CD is represented with inferences in Figure 3. Without necessary variable 'co treatment' and 'side effects' are seen as covariates that do not bias the true causal effect, but as covariates that must not be selected for adjustment (mediator) which may increase bias. Even with necessary variables specification 'side effects' covariate requires adequate reasoning to be considered as a potential candidate for adjustment. Without sufficient cause specification, the covariate 'comorbidity' is seen as a potential candidate for adjustment whereas this adjustment cannot correct bias.

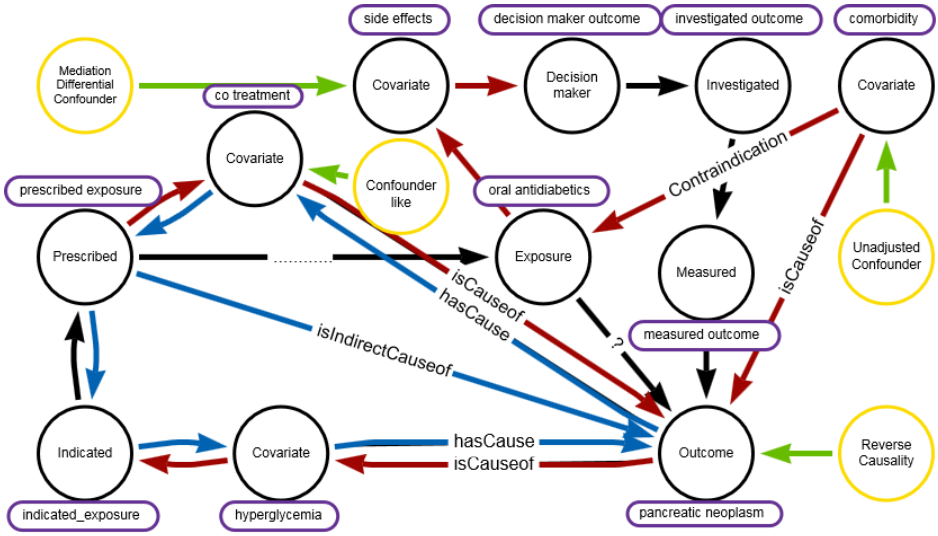

Figure 2. Ontological causal diagram built with OntoBioStat using inferences. Instances name are in purple. Object properties stated are in red, inferred in blue, being a part of automatic construction black. Inferred classes are in yellow.

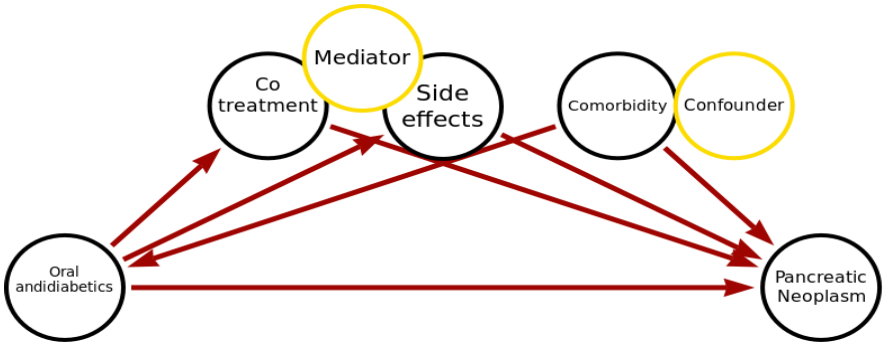

Figure 3. Standard causal diagram with inferences.

4. Discussion and Conclusions

In this article, we showed the usefulness of a novel model following the footsteps of the Directed Acyclic Graph (DAG), CD, and Sufficient Component Cause model [11] that could be used to enhance the consciousness of the study biases. Furthermore, the

pre-existent generic instances provide a significant added value to the knowledge representation of a given study and should help users to reflect on their own practices. Since the aim of OntoBioStat is to support covariate selection, it relies on a sufficient formalism. For example, it does not include distinction between state, event and process or between 'allow', 'maintain', 'perpetuate' relations as defined in the ontology of causal relations [12].

Decision-support systems such as dagitty [13,14] for DAGs provide an easy to use interface. The R package dagitty enables users to specify CD's structure and to obtain the right set of covariates (minimal and sufficient), instrumental variable, and path analysis. However, dagitty does not provide an automatic DAG construction, adapted reasoning for necessary or sufficient cause, nor rich explanation of the results. Actually, OntoBioStat may be seen more as an educational tool or a safety net provider for unskilled biostatistics users than a real decision-support system to be used on daily basis by expert users for two main reasons: (i) reasoning based on rules and axioms do not provide minimal sufficient set of covariates but put forward all covariates that could bias the results, hence minimal set have to be selected manually, (ii) biostatisticians are not familiar with the Protégé.

Directions for future research include: (i) the implementation of OntoBioStat as an operational system named MetBRaYN [7], combining the strengths of dagitty and OBS with an R interface; (ii) the mapping of ontologies object properties with OntoBioStat causal object properties in order to feed with several instances the ontology.

References

[1] Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet. 2002 Jan;359(9302):248–52.
[2] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan;10(1):37–48.
[3] Vanderweele TJ, Robins JM. Empirical and counterfactual conditions for sufficient cause interactions. Biometrika. 2008 Jan 31;95(1):49–61.
[4] Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. Journal of Clinical Epidemiology. 2021 Aug;S0895435621002407.
[5] Smith B, Ceusters W, et al. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46.
[6] Kahn CE. Transitive closure of subsumption and causal relations in a large ontology of radiological diagnosis. Journal of Biomedical Informatics. 2016 Jun;61:27–33.
[7] Pressat Laffouilhère T, Grosjean J, et al. Ontological Models Supporting Covariates Selection in Observational Studies. Stud Health Technol Inform. 2021 May 27;281:1095–6.
[8] Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), 2015 Jun.
[9] Sirin E, Parsia B, et al. Pellet: A practical OWL-DL reasoner. Web Semantics. 2007 Jun;5(2):51-53.
[10] Pearl J. Causality: models, reasoning, and inference. Cambridge, U.K. ; New York: Cambridge University Press; 2000. 384 p.
[11] Rothman KJ, Greenland S. Causation and Causal Inference in Epidemiology. Am J Public Health. 2005 Jul;95(S1):S144–50.
[12] Galton A. States, Processes and Events, and the Ontology of Causal Relations. Frontiers in Artificial Intelligence and Applications. Volume 239: Formal Ontology in Information Systems. 279–292.
[13] Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GTH. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int J Epidemiol. 2017 Jan 15;dyw341.
[14] Ankan A, Wortel IMN, Textor J. Testing Graphical Causal Models Using the R Package "dagitty". Current Protocols. 2021 Feb.