Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210361

Ontological Models Supporting Covariates Selection in Observational Studies

Thibaut PRESSAT LAFFOUILHÈRE[a,b,c,1], Julien GROSJEAN[a,d], Jacques BÉNICHOU[b], 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 EA 4108, F-76000 Rouen, France*
[d] *LIMICS U1142, Sorbonne Université, Paris, France*

Abstract. In the context of causal inference, biostatisticians use causal diagrams to select covariates in order to build multivariate models. These diagrams represent datasets variables and their relations but have some limitations (representing interactions, bidirectional causal relations). The MetBrAYN project aims at building an ontological-based process to tackle these issues. The knowledge acquired by the biostatistician during a methodological consultation for a research question will be represented in a general ontology. In order to aggregate various forms of knowledge the ontology will act as a wrapper. Ontology-based causal diagrams will be semi-automatically built. Founded on inference rules, the global system will help biostatisticians to curate it and to visualize recommended covariates for their research question.

Keywords. causality, covariates selection, ontologies, biostatistics, causal diagram

1. Introduction

Causal diagrams (CDs) [1] are formal knowledge representations used in causal inference by biostatisticians. CD enable to select covariates for biostatistical model in order to estimate causal effect without bias. They have some limitations such as interactions or bidirectional causal relation that lead to use informal tricks [2]. Ontologies, formal representations of knowledge, could help to list and standardize information needed to build a complete CD. Upper domain ontology could be used to produce use case ontologies that outperformed CD in terms of amount of contents and features. These use case ontologies would support biostatisticians in their covariates selection task.

2. Methods

All relevant information collected during biostatistical consultation will be represented in the upper domain ontology "OntoBioStat" (OBS) with a minimal number of classes and relations. OBS will be mapped [3] with other ontologies related to biomedical research, such as the Ontology for Biomedical Investigation [4], and more upper ontologies

1 Corresponding Author, 37 Boulevard Gambetta, 76000 Rouen, France , E-mail: t.pressat@chu-rouen.fr

representing general concepts (e.g. causality) [5]. OBS will intend to semi-automatic aggregate various forms of knowledge representations (ontologies, knowledge graphs, causal diagrams) but also datasets information (variables and metadata). This ontology, instantiated with specific datasets, will allow to build a Causal Diagram Ontology (CDO) that will outperform usual CDs.

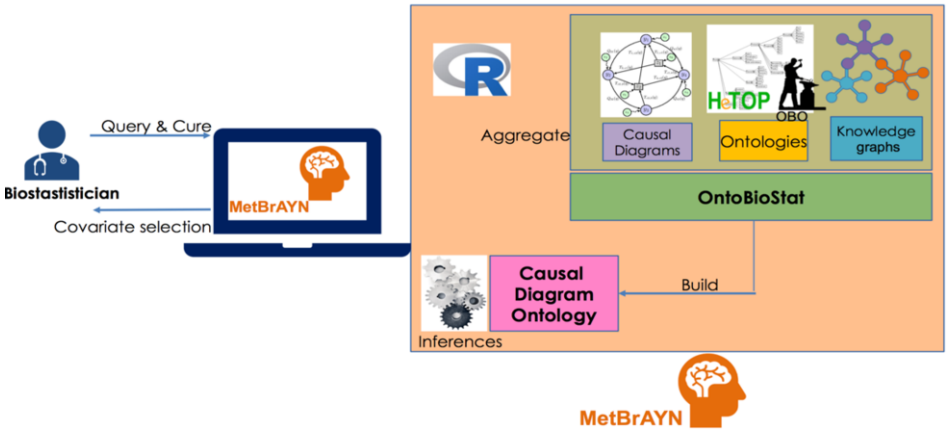

Figure 1. Workflow overview of MetBrAYN.

The biostatistician will be able to query the global system (Fig.1) called "**Met**hodologist **Br**ain is **A**ll **Y**ou **N**eed" (MetBrAYN) for CDO curation. Moreover, thanks to inference rules [6], MetBrAYN will recommend covariates for model building, depending on the initial research question.

3. Expected results and Conclusion

The biostatistician will define (i) the couple (outcome; exposure of interest). For example (Death, Treatment); (ii) the "baseline" (e.g. entry in critical care); (iii) the environment in which the treatment is delivered (e.g. critical care unit); and (iv) the population (e.g. out-of-hospital cardiac arrest). Then, *MetBrAYN* relying on CDO and inference rules will provide: indispensable covariates, optional covariates, prohibited covariates, recoding (e.g. age recode in quintiles), interaction term, sensitivity analysis (e.g. slight definition of treatment), with explanations (e.g. asystolic status was indispensable covariates because it is a confounding variable), and a comment on extrapolation. In this project, the system *MetBrAYN*, that includes OBS and CDO, will avoid modeling mistakes and would be used to train unskilled biostatisticians.

References

[1] Greenland S, et al. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan;10(1):37–48.
[2] Suzuki E, et al. Causal Diagrams: Pitfalls and Tips. J Epidemiol. 2020 Apr 5;30(4):153–62.
[3] Mary M, et al. Ontological Representation of Laboratory Test Observables: Challenges and Perspectives in the SNOMED CT Observable Entity Model Adoption. AIME17, 2017;14-23.
[4] Bandrowski A, et al. The Ontology for Biomedical Investigations. PLoS ONE. 2016 Apr 29;11(4):e0154556.
[5] Besnard Ph, et al. Ontology-based inference for causal explanation1. ICA. 2008 Jul 31;15(4):351–67.
[6] Lamy JB, Soualmia LF. Formalization of the semantics of iconic languages: An ontology-based method and four semantic-powered applications. Knowledge-Based Systems. 2017 Nov;135:159–79.