Modèle ontologique formel, un appui à la sélection des variables pour la construction des modèles multivariés.

Thibaut Pressat-Laffouilhère

Normandie Université

UNIVERSITÉ DE ROUEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **INFORMATIQUE**

Préparée au sein de l'**Université de Rouen Normandie**

Modèle ontologique formel, un appui à la sélection des variables pour la construction des modèles multivariés.

Présentée et soutenue par

THIBAUT PRESSAT-LAFFOUILHERE

Thèse soutenue le 19/12/2023

devant le jury composé de :

MME KAREN LEFFONDRE	Professeur des Universités - UNIVERSITE BORDEAUX 1 SCIENCES ET TECHNOLOGIE	Rapporteur du jury
MME FLEUR MOUGIN	Professeur des Universités - UNIVERSITE BORDEAUX 1 SCIENCES ET TECHNOLOGIE	Rapporteur du jury
M. JEAN CHARLET	Chargé de Recherche - SORBONNE UNIVERSITE	Membre du jury
M. JULIEN GROSJEAN	- Université de Rouen Normandie	Membre du jury
MME CATHERINE DUCLOS	Professeur des Univ - Prat Hospitalier - UNIVERSITE PARIS 13 PARIS-NORD	Président du jury
MME LINA SOUALMIA	Professeur des Universités - Université de Rouen Normandie	Directeur de thèse
M. JACQUES BENICHOU	Professeur des Univ - Prat Hospitalier - Université de Rouen Normandie	Co-directeur de thèse

Thèse dirigée par **LINA SOUALMIA** (LABORATOIRE D'INFORMATIQUE DE TRAITEMENT DE L'INFORMATION ET DES SYSTEMES) et **JACQUES BENICHOU** (Université de Rouen Normandie)

MIIS

litis

Mind over data

Judea Pearl

Modèle ontologique formel, un appui à la sélection des variables pour la construction des modèles multivariés

Résumé

Répondre à une question de recherche causale dans un contexte d'étude observationnelle nécessite de sélectionner des variables de confusion. Leur intégration dans un modèle multivarié en tant que co-variables permet de diminuer le biais dans l'estimation de l'effet causal de l'exposition sur le critère de jugement. Leur identification est réalisée grâce à des diagrammes causaux (DCs) ou des graphes orientés acycliques. Ces représentations, composées de nœuds et d'arcs orientés, permettent d'éviter la sélection de variables qui augmenteraient le biais, comme les variables de médiation et de collision. Les méthodes existantes de construction de DCs manquent cependant de systématisme et leur représentation de formalisme, d'expressivité et de complétude.

Afin de proposer un cadre de construction formel et complet capable de représenter toutes les informations nécessaires à la sélection des variables sur un DC enrichi, d'analyser ce DC et surtout d'expliquer les résultats de cette analyse, nous avons proposé d'utiliser un modèle ontologique enrichi de règles d'inférences.

Un modèle ontologique permet notamment de représenter les connaissances sous la forme de graphe expressif et formel composé de classes et de relations similaires aux nœuds et arcs des DCs.

Nous avons développé l'ontologie OntoBioStat (OBS) à partir d'une liste de questions de compétence liée à la sélection des variables et de l'analyse de la littérature scientifique relative aux DCs et aux ontologies. Le cadre de construction d'OBS est plus riche que celui d'un DC, intégrant des éléments implicites tels que les causes nécessaires, contextuels d'une étude, sur l'incertitude de la connaissance et sur la qualité du jeu de données correspondant.

Afin d'évaluer l'apport d'OBS, nous l'avons utilisée pour représenter les variables d'une étude observationnelle publiée et avons confronté ses conclusions à celle d'un DC. OBS a permis d'identifier de nouvelles variables de confusion grâce au cadre de construction différent des DCs et aux axiomes et règles d'inférence. OBS a également été utilisée pour représenter une étude rétrospective en cours d'analyse : le modèle a permis d'expliquer dans un premier temps les corrélations statistiques retrouvées entre les variables de l'étude puis de mettre en évidence les potentielles variables de confusion et leurs éventuels substituts ("proxys"). Les informations sur la qualité des données et l'incertitude des relations causales ont quant à elles facilité la proposition des analyses de sensibilité, augmentant la robustesse de la conclusion de l'étude. Enfin, les inférences ont été expliquées grâce aux capacités de raisonnement offertes par le formalisme de représentation d'OBS.
À terme OBS sera intégrée dans des outils d'analyse statistique afin de bénéficier des bibliothèques existantes pour la sélection des variables et de permettre son utilisation par les épidémiologistes et les biostatisticiens.

Mots clé : Ontologie, Intelligence Artificielle, Diagramme Causaux, Graphe Orienté Acyclique, Causalité, Médecine, Sélection des variables, Facteur de confusion

Formal ontological model, a support for variable selection in the construction of multivariate models

Abstract

Responding to a causal research question in the context of observational studies requires the selection of confounding variables. Integrating them into a multivariate model as co-variables helps reduce bias in estimating the true causal effect of exposure on the outcome. Identification is achieved through causal diagrams (CDs) or directed acyclic graphs (DAGs). These representations, composed of nodes and directed arcs, prevent the selection of variables that would introduce bias, such as mediating and colliding variables. However, existing methods for constructing CDs lack systematic approaches and exhibit limitations in terms of formalism, expressiveness, and completeness.

To offer a formal and comprehensive framework capable of representing all necessary information for variable selection on an enriched CD, analyzing this CD, and, most importantly, explaining the analysis results, we propose utilizing an ontological model enriched with inference rules.

An ontological model allows for representing knowledge in the form of an expressive and formal graph consisting of classes and relations similar to the nodes and arcs of CDs.

We developed the OntoBioStat (OBS) ontology based on a list of competency questions about variable selection and an analysis of scientific literature on CDs and ontologies. The construction framework of OBS is richer than that of a CD, incorporating implicit elements like necessary causes, study context, uncertainty in knowledge, and data quality.

To evaluate the contribution of OBS, we used it to represent variables from a published observational study and compared its conclusions with those of a CD. OBS identified new confounding variables due to its different construction framework and the axioms and inference rules. OBS was also used to represent an ongoing retrospective study analysis. The model explained statistical correlations found between study variables and highlighted potential confounding variables and their possible substitutes (proxies). Information on data quality and causal relation uncertainty facilitated proposing sensitivity analyses, enhancing the study's conclusion robustness. Finally, inferences were explained through the reasoning capabilities provided by OBS's formal representation.

Ultimately, OBS will be integrated into statistical analysis tools to leverage existing libraries for variable selection, making it accessible to epidemiologists and biostatisticians.

Key words : Ontology, Artificial Intelligence, Causal Diagrams, Directed Acyclic Graph, Causality, Medicine, Variable Selection, Confounding Factor

Remerciements

Je remercie le Pr Lina Soualmia et le Pr Jacques Bénichou d'avoir accepté de codiriger cette thèse.
Je remercie le Dr Julien Grosjean d'avoir été mon encadrant sur place.
Je remercie les membres du jury d'avoir accepté d'évaluer ma thèse.
Je remercie les équipes de biostatistique et d'informatique médicale pour leur bonne humeur et leur bienveillance
Je remercie ma famille, mes amis et ma compagne.
Je dédicace cette thèse à mes deux filles nées pendant sa réalisation

Table des matières

Index des figures

Index des tableaux

Introduction

Représentation des connaissances et sélection de variables

Avec l'apparition des systèmes d'information de taille de plus en plus importante, le nombre d'études sur des données de vie réelle est en constante hausse. Répondre aux questions de causalité avec ce type de données relève d'un véritable défi qui nécessite, entre autres, de corriger les biais via une sélection appropriée de variables. La problématique de sélection de variables concerne de nombreux acteurs, que ce soit en biostatistique ou en informatique (« feature selection »). C'est un défi régulier pour les professionnels qui manipulent des données. Les approches de sélection de variables sont très documentées dans la littérature. La sélection de variables fondée sur une représentation des connaissances n'est pas une méthode qui rencontre un franc succès auprès des biostatisticiens, cependant c'est celle qui offre le moins de biais. Les représentations des connaissances proposées en épidémiologie, comme les diagrammes causaux (Greenland S et al., 1999), sont très utiles pour sélectionner les variables. Ces représentations sont cependant incomplètes car le processus de sélection des variables doit prendre en compte une quantité importante d'informations. Une représentation des connaissances fondée sur un modèle d'ontologie (Bock et al.) offre un cadre de représentation expressif, formel, et standard. Ce canevas permet de représenter les connaissances sous la forme d'un graphe plus riche qu'un diagramme causal. L'ontologie est une représentation intelligente car des inférences logiques peuvent être faites à partir d'axiomes et de règles. Le raisonnement qui a servi à faire ces inférences est transparent, il est donc facile d'expliquer un résultat fourni par l'ontologie.

Dans cette thèse, je présente OntoBioStat, une ontologie que j'ai développée pour aider à la construction, à l'analyse et à la compréhension des diagrammes causaux enrichis afin d'aider le professionnel dans sa tâche de sélection des variables avec pour objectif l'inférence causale. Je présente également plusieurs cas d'étude permettant de valider cette ontologie et son utilisabilité en présence de données de vie réelle.

Contexte de recherche

Les travaux présentés ici traitent de la problématique de la **représentation des connaissances** dans le but de **sélectionner des variables** pour la création de modèles multivariés afin de répondre à une **question de causalité** dans le cadre **d'études observationnelles en recherche biomédicale**.

Durant mon internat de santé publique, je me suis tourné vers la discipline biostatistique et pendant mon assistanat vers l'informatique médicale. J'ai pu participer à de nombreuses études qui utilisent souvent des données de vie réelle. La sélection des variables était un sujet récurrent. J'ai commencé à m'intéresser au sujet de la causalité et des méthodes de sélection des variables correspondantes ainsi qu'aux pratiques des chercheurs. Suite à différents constats sur ce sujet (e.g., méthodes variées, centre dépendantes, erronées, non reproductibles, superficielles, impactant gravement les résultats dans certains cas, etc.), j'ai entrepris cette thèse de sciences dans le service d'informatique médicale du CHU de Rouen au début de mon assistanat.

Les différents objectifs sous-jacents sont les suivants :

- Représenter toutes les informations disponibles qui seraient nécessaires à la sélection de variables ;

- Proposer un cadre formel suffisant regroupant les éléments nécessaires pour aider à la construction d'un diagramme causal ;

- Analyser en permettant une automatisation (ou semi-automatisation) de l'analyse des diagrammes causaux ;

- Expliquer grâce à l'ontologie les résultats inférés.

Tous ces objectifs se réalisent dans un système expert théorisé afin que de futurs développements soient possibles.

Organisation du mémoire

Le Chapitre 1 « Recherche biomédicale, causalité et question de recherche causale » rappelle ce qu'est la recherche médicale et le rôle du biostatisticien, la causalité, les biais et les situations dans lesquelles la sélection des variables peut corriger les biais. Il permet de comprendre dans quels contextes l'ontologie pourrait être utilisée.

Le Chapitre 2 « Méthodes de sélection des variables dans l'inférence causale : état de l'art » définit la sélection des variables dans sa globalité, la sélection guidée par les données, puis la sélection basée sur une représentation des connaissances que sont les diagrammes causaux. Il permet de mettre en lumière les défauts de l'existant que l'ontologie pourrait corriger.

Le Chapitre 3 « Ontologies et sélection des variables dans l'inférence causale dans la recherche biomédicale : état de l'art » définit ce qu'est une ontologie et fait un état des lieux

des ontologies en rapport avec le sujet de cette thèse. Avant d'entamer le cœur du projet, ce chapitre contient quelques rappels et fait la liste des ontologies similaires à celle qui va être construite.

Le Chapitre 4 « Développement d'OntoBioStat : une ontologie pour aider à la sélection des variables dans l'inférence causale » présente la méthodologie de construction de l'ontologie.

Le Chapitre 5 « Expérimentations et cas d'usage d'OntoBioStat » détaille le processus d'utilisation de l'ontologie développée ainsi que les trois cas d'usages réalisés avec l'ontologie.

Enfin, la partie « Conclusion et perspectives » permet de dresser un bilan et de proposer des pistes d'amélioration de ces travaux de thèse.

Chapitre 1 : Recherche biomédicale, causalité et question de recherche causale

1. La recherche biomédicale et le méthodologiste biostatisticien

La recherche biomédicale regroupe l'étude d'entités en rapport avec la médecine et la compréhension du corps humain. Ces entités peuvent être une maladie, un traitement, ou encore un gène. L'article 'L. 1121-1' du code de la Santé Publique définit la recherche biomédicale comme « ***Les recherches organisées et pratiquées sur l'être humain en vue du développement des connaissances biologiques ou médicales*** ». Cette définition regroupe les recherches dites *interventionnelles*, qui comportent une intervention en dehors du cadre du soin courant (c-à-d. le soin qui se déroule comme habituellement) et *non interventionnelles* (c-à-d. observationnelles) dans lesquelles tous les actes font partie du soin courant. Cette définition inclut uniquement les études non interventionnelles (observationnelles), dites *prospectives*, avec un protocole de recherche réalisé en amont de l'étude. Des études dites *rétrospectives,* parfois sans protocole particulier, font pourtant partie de la recherche biomédicale. Il existe par exemple les études réalisées à partir de données provenant des bases médico-administratives (e.g., le Système National des Données de Santé en France) ou les entrepôts de données de santé des établissements de santé. Ces données, collectées avec pour but premier la facturation des soins ou la gestion des soins courants, sont aussi appelées *données de vie réelle*.

En recherche biomédicale, le rôle du biostatisticien peut être défini de la façon suivante : « *Biostatisticians play an important role for the planning and conduct of a study,*

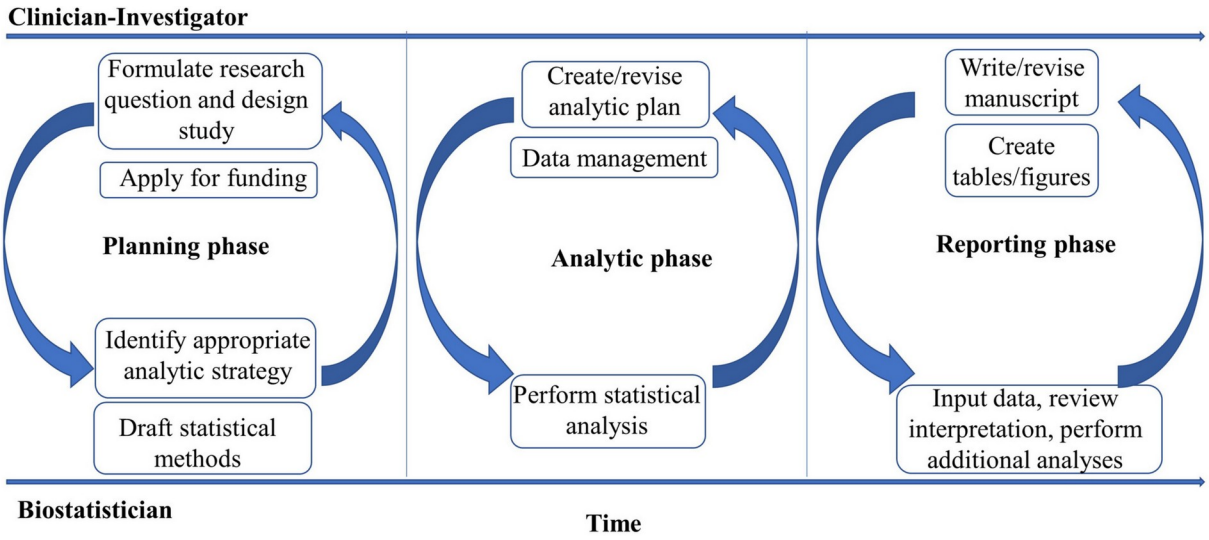


Figure 1. Activités du biostatisticien au cours de la recherche. Zapf A et al 2019

for the analysis of the data, and for the reporting of the results » (Zapf A et al 2019, Lee J et al 2022).

De manière plus précise, le méthodologiste biostatisticien va fournir une expertise dans :

- la réalisation du protocole : (i) *la formulation de la question*, (ii) *la définition de(s) intervention(s) et des critères de jugement*, (iii) *les critères d'inclusion et exclusion*, et (iv) *les statistiques* (c.à-d. le nombre de sujets nécessaires, les méthodes d'analyses classiques ou plus novatrices) ;
- le contrôle du formulaire de saisie e(*case report form*) et la cohérence des données parfois conjointement avec un gestionnaire des données (*data manager*) ;
- l'analyse statistique ;
- le rapport d'analyse, l'interprétation et l'explication des résultats ;
- l'écriture et la relecture du manuscrit jusqu'à publication.

Durant l'analyse statistique, le biostatisticien manipule les données de patients sous la forme de variables (ou *features*). Dans le Larousse, on définit une variable par une « *grandeur susceptible de varier dans un ensemble donné, et telle qu'à chaque valeur prise par cette grandeur puisse correspondre, au moins théoriquement, un effectif de personnes ou une fréquence en pourcentage* ». La variable est à différencier de la constante qui elle ne varie pas dans un ensemble donné comme la situation géographique d'un hôpital.

La façon de traiter les données dépend grandement de l'objectif de la recherche. Ici, nous nous intéresserons tout particulièrement aux objectifs liés à la causalité.

2. Causalité en recherche biomédicale

En recherche biomédicale, la *causalité* au niveau **populationnel** repose sur une définition **probabiliste** et **contrefactuelle (**Lewis D et al 1973, Pearl J 2001).

Niveau Populationnel. La causalité est prouvée à un niveau populationnel et non individuel. Si un effet causal est mis en évidence, c'est pour un *individu moyen*, ou autrement dit en moyenne. Cet effet causal n'est pas vrai pour tous les individus. Par exemple, il a été prouvé que fumer du tabac cause le cancer du poumon ; or, face à patient fumeur avec un cancer du poumon, on ne pourra jamais affirmer avec certitude si, pour ce patient, le cancer du poumon est dû au tabac ou à une autre exposition. De la même manière, nous savons qu'un traitement A est meilleur qu'un traitement B. Nous proposerons le traitement A à un patient donné, même si nous ne sommes pas sûrs qu'il soit meilleur pour ce patient en particulier ; ce qui s'oppose à un paradigme appelé « médecine personnalisée » qui vise justement à s'adapter à l'individu et pas seulement à une population dans laquelle la causalité a pu être prouvée statistiquement.

Définition Probabiliste. La variable A est la cause de la variable B si et seulement si l'occurrence de la variable A augmente la probabilité d'occurrence de la variable B. En recherche biomédicale, les effets déterministes en opposition avec les effets probabilistes (ou stochastiques) sont rares, d'où l'intérêt de s'appuyer sur la probabilité d'occurrence d'un évènement. En effet, en reprenant l'exemple précédent, le tabac n'entraîne pas chez tout le monde un cancer du poumon. En revanche, il augmente la probabilité d'occurrence de celui-ci par rapport à un non fumeur. Cette seule définition est trop souple car l'occurrence conjointe de deux variables (corrélation) est possible pour d'autres raisons qu'une raison de causalité. En effet, à l'échelle populationnelle la taille du bras et la taille de la jambe d'un individu sont corrélées mais la taille de l'une ne cause pas l'autre. Néanmoins, elles partagent des causes communes comme la génétique et l'environnement.

Définition Contrefactuelle. En philosophie, la définition du terme « contrefactuel » est : « *Qui renvoie à la réflexion sur les événements qui ne se sont pas réalisés mais auraient pu le faire sous certaines conditions* » (Lewis D et al 1973). En résumé, ce sont des évènements qui sont contraires aux faits. Ainsi, il existe un lien causal entre A et B lorsque nous répondons par l'affirmative à l'énoncé suivant : « si A n'avait pas eu lieu, alors B n'aurait pas eu lieu ». Soit par exemple, un groupe de patients qui subissent une chirurgie pour guérir d'un cancer. En se plaçant dans une situation contrefactuelle, il faut imaginer ce qu'il se serait passé si les patients n'avaient pas subi l'intervention chirurgicale. C'est pour cela qu'en recherche, nous

avons besoin d'un second groupe de patients « *similaires* » aux précédents pour simuler ce qu'il se serait passé.

Le mot *causal* se distingue des mots « association » ou « associés ». Les études qui utilisent le mot *association* sont courantes. Ce terme d'association englobe tout ce qui est causal vrai direct ou indirect mais aussi tout ce qui est simplement 'corrélé' statistiquement. La peur d'utiliser le mot causal est réelle dans la communauté scientifique. Si ce n'est pas la peur qui saisit la communauté, c'est un vieux dogme qui la retient : « *seul les essais cliniques randomisés répondent à une question causale* ». Par exemple, le Journal of the American Medical Association (JAMA), un des grands journaux médicaux, utilise le mot *association* dans 42/231 des titres des articles publiés en 2021 et le mot *effect* (qui renvoie au mot *cause*) dans 80/231 titres. Les 42 concernent des études observationnelles, et les 80 sont des essais cliniques randomisés. Cela entraîne beaucoup d'approximations sémantiques (e.g., association), une souplesse méthodologique délétère et une interprétation faussée des résultats (Hernán MA 2018). Il est donc important de bien définir ce qu'est une question de recherche causale.

3. Question de recherche causale

3.1. Objectifs causaux

Une question de recherche causale (ou étiologique) peut correspondre à deux objectifs :

(i) évaluer l'effet d'une exposition ou d'un facteur modifiable qui peut être un comportement (e.g., comportement sexuel à risque, utilisation de seringues non stériles, consommation de tabac), un environnement (e.g., la pollution de l'air), une intervention extérieure (e.g., traitement antirétroviraux), une maladie (e.g., diabète, hypertension) sur un critère de jugement ;

(ii) explorer le côté plus « mécanistique » en étudiant parfois ce qu'on appelle des facteurs ou expositions non modifiables telles que la génétique, ou le sexe biologique.

Le premier objectif permet d'aider à la prise de décision médicale lorsque par exemple le médecin a le choix entre deux traitements. Grâce à l'étude, le médecin a une estimation de ce qu'il se passerait sous le traitement A par rapport au traitement B en moyenne (Hernán M et al 2021). Le médecin choisira le traitement qui, en moyenne, est meilleur. En résumé, répondre à des questions de recherche causale permet d'agir sur les expositions ou facteurs modifiables. Dans notre exemple, si la consommation de tabac a un effet causal délétère sur la santé au niveau populationnel, il est préférable d'arrêter la consommation de tabac.

Le deuxième objectif permet d'enrichir la connaissance des liens causaux entre entités.

3.2. Les autres types de questions

Il arrive que différents types de questions de recherche soient posés dans la même étude, sur les mêmes données. Cependant, les méthodes pour y répondre et l'interprétation des résultats sont très différentes. Aujourd'hui encore, il n'est pas facile lors de la lecture d'articles scientifiques de comprendre clairement le type de question de recherche (causale versus prédictive), ce qui participe à la confusion entre les deux types de questions et l'interprétation des résultats. De plus, les auteurs font des erreurs d'interprétation dans un sens comme dans un autre. Par exemple, dans 14/47 des études prédictives analysées par Ramspek CL et al. 2021, les auteurs interprètent les poids des variables comme des effets causaux. Il est cependant nécessaire de distinguer les questions causales des questions de *prédiction*, de *description,* ou d'*exploration* :

(i) Les questions *prédictives* incluent des études diagnostiques et pronostiques dont le but est de déterminer l'état présent ou futur d'un individu à partir d'un faisceau d'arguments (prédicteurs). Ces prédicteurs ne doivent pas être nécessairement des causes. En effet, par exemple, les dents et les doigts jaunis prédisent peut-être très bien le cancer du poumon mais ne sont en aucun cas des causes de celui-ci ;
(ii) Les questions *descriptives* ont pour but d'étudier un phénomène sans préjuger des relations entre variables, comme par exemple l'étude de l'incidence du cancer du sein au cours du temps (en 2015, 2016, 2017, ...etc.).
(iii) Les questions *exploratoires* traitent de sujets complètement inconnus ou insuffisamment connus. C'est la partie biologique de la recherche biomédicale qui étudie différents marqueurs omiques (e.g., radiomique, métabolomique, génomique, protéomique, …etc.). Ces données de hautes dimensions sont aussi utilisées pour des études de prédiction.

3.3. Expliciter la question de recherche causale

Une question de recherche peut être formulée de bien des manières mais reste une tâche difficile. Un grand classique est d'utiliser PICO(T) (Richardson WS et al 1995), un acronyme pour (i) Population ou Patient problem, (ii) Intervention, (iii) Comparison and (iv) Outcome, ((v)time). Il existe différentes variantes telles que PEO : (i) Population (ii), Exposure, and (iii) Outcome (Kestenbaum B et al 2018). Les termes employés « exposition » et « critère de jugement » sont fortement utilisés lorsque le but de l'étude est de mettre en évidence un effet causal. Le fait de retrouver le mot exposition plutôt que prédicteur ou

facteur pronostique est un fort indice en faveur de la recherche d'un effet causal. Ces deux méthodes ne font jamais mention de facteur pronostique, diagnostic, incidence, etc. Des utilisateurs peu avertis pourraient utiliser le mot exposition à la place de facteur pronostique et confondre le type de question de recherche.

Nous verrons plus loin que les essais cliniques randomisés répondent à des questions de causalité par essence. Ainsi, pour Hernan, lorsque l'on utilise une étude observationnelle pour répondre à une question de recherche causale, préciser que nous tentons d'émuler un essai hypothétique (« the target trial ») est une manière d'annoncer clairement que le but de la recherche est de mettre en évidence un effet causal (Hernan M et al 2016). Finalement, la question de recherche pourrait être de la forme « Est-ce que l'exposition X cause le critère de jugement Y ? ». Une fois bien formulée, le méthodologiste pourra proposer différents types d'études et méthodes pour y répondre.

4. Comment répondre à une question de recherche causale

4.1. Les essais cliniques randomisés

Les essais cliniques randomisés font partie des études dites *interventionnelles* ou *expérimentales*. Dans un essai clinique randomisé en double aveugle, il y a deux groupes ou deux bras de sujets, traités par deux traitements différents ((A et B) ou (A et Placebo)). L'assignation du traitement est *randomisée* et *secrète*, donc aléatoire et inconnue à l'avance (par le sujet et le médecin). Cette assignation est imposée : c'est une *intervention* de la part du chercheur qui altère la raison pour laquelle un patient reçoit un traitement.

Aujourd'hui, la référence pour prouver l'existence d'un lien causal entre exposition (e.g., chimiothérapie versus chirurgie) et un critère de jugement (e.g., rechute ou décès) est l'*essai clinique randomisé en double aveugle*. En effet, dans la pyramide des preuves, il est au plus haut avec les méta-analyses. Cependant, la population ne reflète pas toujours celle chez qui ont va appliquer le traitement (comme le vaccin contre la COVID-19 chez les patients âgés de plus de 75 ans). De plus, il n'est pas toujours évident de réaliser de telles études pour des raisons éthiques. En partant du postulat (très fort) que seul un essai clinique randomisé peut répondre à une question de causalité sans biais, il faudrait assigner à des sujets de manière aléatoire l'exposition « tabac » pour prouver l'effet causal du tabac sur le cancer du poumon. On ne peut décemment pas faire un essai clinique randomisé dans cette situation qui n'est qu'un exemple parmi tant d'autres. Comme l'affirme Rutter M. *« That cannot be a solution,..., for a mixture of both ethical and practical reasons » Rutter M 2009*.

4.2. Les études observationnelles

Une étude observationnelle est une étude dans laquelle le chercheur se place en tant qu'observateur de l'effet d'une exposition (e.g., traitement, comportement de santé, ...etc.). Contrairement aux essais cliniques randomisés, l'exposition et le suivi vont dépendre à la fois des caractéristiques du sujet et du médecin qui le prend en charge. L'étude observationnelle peut être réalisée à partir de données de soins courants déjà disponibles dans les dossiers informatisés ou des entrepôts de données de santé, ou en suivant les patients de manière prospective (toujours sans intervenir sur l'exposition). Le chercheur peut être amené à enquêter pour compléter les données dont il dispose en interrogeant les patients ou en analysant du matériel sanguin conservé par exemple.

4.2.1. Les schémas d'étude (« designs »)

Il existe différents « designs » d'études observationnelles tels que cohorte, cas témoin, et transversales pour les plus connus d'entre eux. Tous les designs d'études observationnelles ne peuvent pas être utilisées pour répondre à une question de causalité :

(i) Les études transversales étudient les caractéristiques d'un patient à un instant donné. Il n'existe aucune notion chronologique entre variables. C'est à dire qu'il n'est pas possible de savoir si A est apparue avant B ou l'inverse. De ce fait, on ne peut affirmer avec précision si le patient était exposé avant que sa maladie ne se déclare ;

(ii) Les études écologiques dans lesquelles l'unité statistique ne peut pas être le patient ou l'individu. C'est à dire que l'information ou les données ne sont pas collectées par patients mais par lieu ou fenêtre temporelle sans moyen de revenir à l'unité « patient ». Par exemple, elles mesurent des corrélations entre consommation de beurre et incidence d'infarctus mois par mois en France.

4.2.2. Les biais retrouvés dans les études observationnelles

Le vrai effet causal est quelque chose d'abstrait, d'inatteignable qui devient visible lorsqu'il traverse le prisme de la réalité. Nous ne sommes en capacité que de voir le vrai effet causal déformé. Ces déformations sont appelées *biais* (« deviation of the truth » Grimes DA et al 2002) et amènent à deux issues possibles :

(i) Aucun effet. Certaines déformations sont *systématiques* quelque soit le groupe d'exposition (comme par exemple une balance qui se trompe toujours de deux kilogrammes). Ce sont les biais dits *non différentiels*, car ils ne diffèrent pas selon le groupe d'exposition des patients. Lorsque l'on estime la différence entre le groupe exposé et le groupe non exposé les biais s'annulent. Ils ne biaisent donc pas l'estimation du vrai effet causal ;

(ii) Estimation biaisée du vrai effet causal vers le haut ou vers le bas. Contrairement aux précédents ce sont des biais *différentiels.* Les biais sont parfois si importants qu'on observe un lien causal qui n'existe pas réalité ou inversement, ou que l'effet observé a un sens inverse par rapport à la réalité. En effet, dans le classique paradoxe de Simpson, l'effet d'un traitement A observé peut apparaître comme supérieur au traitement B dans la population alors que chez les hommes, c'est le traitement B qui est supérieur à A, et chez les femmes c'est aussi le traitement B (Julious SA et al 1994).

Même si les essais cliniques randomisés peuvent être biaisés, ils sont constitués de telle sorte que les biais soient minimes. Pour chaque type de biais, je rappellerai la méthode utilisée dans les essais cliniques pour l'éviter, puis définirai le biais. Il existe trois grands types de biais : ***confusion***, ***mesure*** *et* ***sélection*** *:*

Confusion. La différence fondamentale entre une étude observationnelle et un essai clinique randomisé est l'absence de randomisation. La randomisation du traitement effectuée dans les essais cliniques randomisés est la clé du contrôle *des biais de confusion* (Armitage P. 2003). Reprenons l'exemple de la chirurgie versus la chimiothérapie pour guérir le cancer. Si l'attribution du traitement est aléatoire ou randomisée, les caractéristiques des patients (connues ou non) seront réparties (ou distribuées) de manière équilibrée entre les deux groupes de traitement. Par exemple, il y aura autant d'hommes dans le groupe A que dans le groupe B, ou autant de personnes âgées. Les caractéristiques des patients (connues ou non) qui influencent la réponse au traitement, seront donc elles aussi identiquement distribuées dans les deux bras. Si les deux groupes de patients sont comparables sauf sur le traitement reçu (chimiothérapie ou chirurgie), alors si la mortalité diffère entre les groupes, c'est bien le traitement qui doit en être responsable. L'assignation aléatoire est donc la seule manière d'éviter l'obtention d'une estimation de l'effet biaisée due aux caractéristiques (c-à-d. variables). Dans une étude observationnelle, l'attribution d'un traitement (e.g., chirurgie, chimiothérapie) ou un comportement (e.g., tabac, alcool) n'est plus aléatoire et dépend des caractéristiques du patient, du médecin, ...etc. Les caractéristiques des patients réparties de manière asymétrique entre les deux groupes risquent de biaiser l'estimation de l'effet causal. Il existe beaucoup de définitions formelles de ce qu'est un biais de confusion ou variable de confusion (variable responsable d'un biais de confusion).

Dans le cadre de nos travaux, nous retiendrons les définitions suivantes : une variable est une variable de confusion si (i) elle cause la maladie ou qu'elle est une variable qui mesure de manière indirecte cette cause de la maladie, dans la population et si (ii) elle est positivement ou négativement corrélée avec l'exposition sans être causée par celle-ci. Une variable est une

variable de confusion si elle cause le critère de jugement, et que cette variable est associée avec l'exposition, mais que cette variable n'est pas un descendant (ou une conséquence) de l'exposition. En d'autres termes, le groupe des exposés diffère des non exposés pour une autre raison que l'effet de l'exposition étudiée (Robins JM et al 1987, Greenland S et al 1986, Miettinen OS et al 1981).

Mesure. Dans un essai clinique randomisé, grâce à un protocole identique pour tous les médecins et le fait que l'essai est en double aveugle, le personnel chargé de mesurer les caractéristiques des patients le fera toujours de la même manière. Dans les études observationnelles, lorsque un médecin recueille les informations par téléphone pour les exposés et via un auto-questionnaire pour les non exposés cela peut entraîner un biais de mesure. Dans le cas d'études cas-témoins, le principe est de recruter des personnes déjà malades (e.g., cancer du poumon) et des non malades puis de leur poser des questions pour savoir s'ils ont été exposés à telle ou telle substance (e.g., tabac). Cette méthode expose à ce que l'on appelle un *biais de mémorisation* correspondant au fait que les personnes malades se souviendront mieux que les non malades. Le biais de mesure aussi appelé biais de classement correspond à une erreur dans la mesure des variables, notamment l'exposition et le critère de jugement. Cela peut entraîner par exemple la classification d'un patient dans le groupe non fumeur alors qu'il fume.

Sélection. Dans un essai clinique, la randomisation du traitement et le suivi en aveugle identique pour tous les patients permettent d'assurer une comparabilité des patients tout au long de l'étude. Néanmoins, le résultat des essais cliniques randomisés peut être biaisé à cause du *biais d'attrition,* les patients sortant de l'étude pour diverses raisons. Lorsque les sorties sont aléatoires cela n'entraîne aucun biais. En revanche, lorsque les sorties sont directement reliées aux effets secondaires éventuels d'un des traitements, ou l'absence d'effet, alors le biais peut être important. Dans les études observationnelles, les caractéristiques des patients déterminent la prescription des médicaments, le protocole appliqué et l'attention portée au patient au cours du temps. Sont parfois regroupés dans biais de sélection, des biais de sélection *à l'inclusion* et les biais de sélection sur *la comparabilité initiale* des groupes et durant le suivi. Concernant les biais d'inclusion, ils peuvent venir de critères de sélection imparfaits. Par exemple, sélectionner les utilisateurs d'un médicament prévalents plutôt que les utilisateurs incidents (i.e., prevalent users *vs.* new users). [Hernan M et al 2004]

En ayant conscience de ces multiples déformations et des différents mécanismes responsables de ceux-ci, il est possible d'estimer le vrai effet causal ou estimer l'amplitude et

le sens du biais. La plupart du temps, il est dit que seuls les biais de confusion peuvent être corrigés (« otherwise internal validity is doomed » : Grimes DA et al 2002). Nous verrons dans la section des diagrammes causaux que cela dépend des sous-types de biais d'information et de sélection, et d'autres éléments, comme le côté 'déterministe' et 'mesurable' de ce qui les a engendrés.

4.2.3. Correction des biais

Certains de ces biais, notamment les biais de confusion, peuvent être corrigés grâce à certains designs telles que les études quasi expérimentales et les *self-controlled design* :

(i) Dans les études quasi-expérimentales, le traitement ne dépend plus des caractéristiques d'un patient mais uniquement d'une fenêtre temporelle, d'un lieu, ou d'un praticien (e.g., étude « ici-ailleurs », « avant-après ») (Bärnighausen T et al 2017). Il peut paraître péremptoire pour certains d'affirmer que les biais de confusion sont corrigés avec une étude quasi-expérimentale. Cependant, dans la théorie, si elle est bien menée, les groupes seront comparables et donc il n'y aura pas de biais de confusion ;

(ii) Dans les études *self-controlled* (*case cross over* et *self controlled case series*), le patient est son propre témoin (Cadarette SM et al 2021).

En l'absence de ces types de designs on peut corriger les biais de confusion soit par ***restriction (exclusion)***, ***appariement***, ***stratification***, ***pondération*** *ou* ***ajustement***.

La restriction consiste à exclure les patients qui possèdent une caractéristique qui biaisent l'effet causal comme, par exemple, exclure les fumeurs pour essayer de mettre en évidence un effet causal entre café et cancer de vessie (fumer étant responsable de l'augmentation du risque de cancer de vessie et étant associé à la prise de café) ;

L'appariement est une méthode qui peut être utilisée au moment de la sélection de la population avant le début de l'étude ou après cela permet d'inclure des patients qui ont des caractéristiques similaires connues dans chaque groupe ;

La stratification consiste à estimer l'effet par strate d'une variable responsable d'un biais puis à rassembler les effets pour obtenir l'estimation finale. Par exemple : l'effet café/cancer de la vessie chez les fumeurs, puis chez les non fumeurs ;

La pondération consiste à pondérer les patients en fonction de la probabilité d'être exposé ou la distribution réelle d'une variable par exemple ;

L'ajustement est une méthode qui consiste à construire un modèle statistique stochastique (appelé modèles multivariés) avec plusieurs variables (covariables) autre que l'exposition responsables de biais, ce qui permet de 'contrôler' le biais. La pondération et l'appariement peuvent aussi être réalisés grâce à un modèle multivarié.

Un modèle statistique est une idéalisation de la réalité fondée sur une distribution théorique d'observations bien réelles. Il ne faut pas confondre les outils de modélisation déterministes, telles que les équations différentielles ordinaires ou équations aux dérivées partielles, et les modèles stochastiques (aléatoires). Ces derniers permettent de décrire les variations aléatoires de données observées. Dans son plus simple apparat, un modèle contient deux variables, X (variable explicative ou indépendante) et Y (à expliquer ou dépendante correspondant au critère de jugement principal). Y = aX + b. En causalité, il y aura le critère de jugement, l'exposition et les autres variables (covariables). Il existe de nombreux modèles, mais les plus utilisés en recherche médicale sont la régression logistique, la régression linéaire et le modèle de Cox.

L'avantage de la méthode d'ajustement, ou des méthodes d'appariement/pondération basées sur des modèles multivariés, est que le nombre de variables prises en compte peut être (très) important. En effet, avec la méthode de restriction, on imagine mal exclure les patients dès qu'ils ont une caractéristique *x* ou *y* ou *z*, de même pour la méthode d'appariement s'il doit y avoir autant de *x*, *y*, *z*, *k*, *l*, *m* dans chaque groupe. Cela réduirait considérablement le nombre de patients inclus, donc la puissance de l'étude, et donc le risque de ne mettre aucun effet causal en évidence alors qu'il pourrait exister. Toutes ces méthodes ont un point commun : elles nécessitent toutes au préalable de ***sélectionner les variables*** qui biaisent ou pourraient biaiser l'effet.

Dans les essais randomisés, les variables d'ajustement, si elles sont incluses dans le modèle, ne serviront qu'à améliorer la puissance de l'essai clinique et à minimiser le biais de l'estimation de la taille d'effet.

Synthèse du chapitre

Dans ce chapitre, il a été question de définir la recherche biomédicale, le rôle du biostatisticien, la causalité, les questions de recherche causale et les études qui permettent de répondre aux questions de recherche causales. Nous avons vu qu'il existait un risque important de se tromper de terme (e.g., association), de question (e.g., prédiction) ou de type d'étude (e.g., transversale). Corriger les biais, dans une situation bien précise (fig de synthèse), via une bonne sélection des variables, pour estimer le vrai effet causal, est le cœur de cette thèse.

Schéma de synthèse

Figure 2. Schéma de synthèse du Chapitre 1.

Bradford Hill appartient-il au passé ?

Il n'est pas permis de terminer ce chapitre sur la causalité sans parler des critères de Bradford Hill. Ils sont encore enseignés, comme étant une source viable pour définir si oui ou non une association statistique retrouvée entre deux variables est bien causale. Ils sont aujourd'hui discutés (Ioannidis JPA 2016, Shimonovich M et al 2021). Pour comprendre d'où viennent ces critères, il faut connaître le contexte. Nous sommes aux Etats Unis dans les années 1950. La guerre fait rage entre industriels du tabac et médecins. Les résultats des études des médecins qui utilisent souvent des méthodes du type cas témoins sont remis en cause par les industriels. Bradford Hill publia alors en 1965 neuf critères qui pourraient aider à statuer sur la nature causale d'une relation statistique. Il résume donc un faisceau d'arguments en neuf points qui sont les suivants : Force d'association (coefficient élevé), Stabilité (répétition dans le temps et l'espace et lieu), Cohérence (« Data should not seriously conflict with the generally known facts of the natural history and biology of the disease »), Spécificité, Temporalité (la cause précède les conséquences), Relation dose effet (gradient biologique), Plausibilité, Preuve expérimentale, Analogie. La force d'association statistique et la stabilité dans le temps de l'association peuvent être retrouvées pour des corrélations simples. Par exemple, la taille des bras et des jambes sont très corrélées (force), cette observation est valide dans le temps et reproductible dans d'autres populations (stabilité) mais elles ne se causent pas l'une l'autre. Inversement, une faible association ne veut pas dire qu'il n'existe pas une relation causale (Rothman KJ et al 2005). La relation dose effet est largement remise en cause en toxicologie moderne avec notamment des effets en U et même en médecine avec par exemple le risque de mortalité et la mesure tensionnelle. Les preuves expérimentales, telles que les résultats d'essais randomisés sont très utiles puisque la randomisation aura supprimé une partie importante des biais. La plausibilité a posteriori des résultats est dangereuse car on peut toujours trouver une explication douteuse à tout. La plausibilité et la cohérence a priori couplées à la temporalité sont très importantes. En effet même si la nuit succède au jour, le jour n'est pas la cause de la nuit. L'analogie est un outil qui sert à imaginer des scénarios en se basant sur des histoires existantes, mais sûrement pas à valider un lien causal (Shimonovich M et al 2021). La situation de spécificité qui correspond au fait qu'il n'existe que l'exposition A qui cause le critère de jugement B et qu'elle ne cause rien d'autre, est rare bien que vraie. SI plusieurs études observationnelles ont tenu compte de (i) la plausibilité biologique a priori, (ii) la cohérence a priori, (iii) la temporalité, (iv) la preuve expérimentale, (v) qu'elles sont

concordantes (stable) et (vi) que la force de l'association est tellement importante qu'une potentielle variable de confusion devrait être vraiment forte (et donc connue) pour annuler le dit effet causal; ALORS oui, il existe une forte chance que l'effet causal soit non nul. Les quatre premiers points regroupent les connaissances, alors que les deux derniers correspondent à l'observation.

Chapitre 2 : Méthodes de sélection des variables dans l'inférence causale : état de l'art

1. Définition de la sélection des variables

La sélection des variables correspond à l'étape de **choix des variables ou covariables** à inclure dans un modèle multivarié. Cette étape va au-delà de la simple inclusion de la variable telle qu'elle a été recueillie à l'origine. Le ***recodage*** d'une variable (de nature quantitative continue par exemple) et l'inclusion d'un terme d'*interaction* entre variables font partie de la problématique de la sélection des variables (Sauerbrei W et al 2020):

Le Recodage. Une variable peut être secondairement créée ou recodée en fonction du contexte. Par exemple, on peut utiliser l'âge en catégories en le découpant en quintile, en probabilité de décès, ou le traiter comme un polynôme âge+âge au carré.

L'Interaction. Les termes d'interaction pourront être utilisés en cas de synergie ou d'antagonisme entre deux variables. Par exemple, pour le risque de survenue de cancer ORL la présence de tabac et alcool chez un même patient augmente le risque de manière synergique.

Faire une erreur dans la sélection des variables peut créer un modèle erroné. Au mieux, il répondra à la question de manière imprécise (par exemple surestimer ou sous-estimer l'effet d'un traitement) au pire il répondra à côté de la question posée. Sachant cet écueil possible, la sélection des variables devra reposer sur des méthodes de sélection adaptées.

2. Les méthodes de sélection des variables

Plusieurs approches ont été développées pour sélectionner les covariables d'un modèle multivarié pour l'inférence causale dans les études observationnelles (Witte J et al 2019). Les deux approches classiques sont d'une part (i) l'utilisation des connaissances préalables sur le sujet, qui est une combinaison d'entretiens avec les cliniciens et de synthèse de la littérature scientifique, et d'autre part (ii) les méthodes dépendantes des données ou guidées par les données (*data-driven*) qui reposent exclusivement sur le contenu du jeu de données fourni pour l'analyse statistique.

En pratique, il arrive régulièrement que ces deux approches soient utilisées conjointement. Le principe et des exemples de méthodes guidées par les données et de méthodes basées sur la connaissance sont présentées puis leurs limites sont exposées.

2.1. La sélection des variables guidée par les données

2.1.1. Méthodes guidées par les données génériques

Il existe deux approches classiques, nommées « basée sur les tests » et « basée sur la pénalisation » dans la revue parue en 2018 réalisée par Desboulets L et al. Toutes ces méthodes sont aussi bien utilisées dans un contexte de causalité que dans un contexte de prédiction (construction de modèles prédictifs).

Approches Basées sur les tests

Ces approches regroupent les procédures pas-à-pas (*stepwise*), pas-à-pas descendant (*backward*) ou pas-à-pas ascendant (*forward*) (Hamaker HC et al 1962). Dans la procédure de pas-à-pas descendant (*backward elimination*), un modèle comprenant toutes les variables pré-sélectionnées est estimé, puis les covariables dont la *p*-valeur est la plus grande sont successivement retirées du modèle jusqu'à ce que toutes les variables restantes soient

significatives à un degré de significativité pré-établi. Dans la procédure de pas-à-pas ascendant (*forward*), un modèle avec seulement la variable d'exposition d'intérêt est estimé, puis les covariables dont la *p*-valeur est la plus petite (conditionnellement à l'exposition) sont ajoutées successivement, une à une, jusqu'à ce que plus aucune nouvelle variable candidate ne soit plus significative, à un seuil de significativité préétabli, lorsqu'elle est intégrée en tant que nouvelle covariable. Le pas-à-pas (*stepwise*) est un mix de ces deux méthodes dans lequel une variable peut être exclue puis incluse dans le modèle à une étape ultérieure. La sélection d'une variable ne se fait pas toujours sur la *p*-valeur de la variable mais sur la différence entre deux modèles. On parle de comparaison de modèles emboîtés (modèle 1 avec *n* variables *vs*. modèle 2 avec *n*+/-1 variables). La qualité prédictive des modèles est d'abord mesurée par exemple, par l'AIC (Critère d'information d'Akaike) ou le BIC (Critère d'information bayésien), avant d'être comparée avec un test (AIC modèle 1 *vs.* AIC modèle 2).

Approches Basées sur la pénalisation

Le deuxième regroupe les méthodes telles que le LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani R 1996) et l'Elastic Net (Zou H et al 2005) qui, contrairement à la pénalisation ridge, permettent de sélectionner des variables. Ces méthodes appliquent une pénalité aux coefficients des variables. Cette pénalité réduit le coefficient de certaines variables à zéro ce qui entraîne l'exclusion de ces variables. Ici, c'est la valeur de la pénalité qui est guidée par les données. La pénalité optimale peut être calculée par validation croisée (e.g., Leave One Out Cross-Validation), ou avec l'AIC. Des variantes de ces méthodes existent, comme le group-lasso (Yuan M et al 2006) qui applique une pénalité différente dans chacun des groupes de variables prédéfinis (Utilisé dans le cas d'une variable catégorielle afin d'appliquer la même pénalité à tous les coefficients de cette variable par exemple).

Le cas de l'apprentissage automatique (ou machine learning) : les méthodes guidées par les données incluent les méthodes de l'apprentissage automatique qui créé automatiquement des modèles à partir d'un certain nombre de variables initiales. L'importance de chaque variable (ou feature) peut être évaluée à la fin et ce résultat, est utilisé afin de conserver uniquement les plus importantes dans le modèle prédictif final. Ces méthodes ne font pas encore partie du quotidien des biostatisticiens, même si certaines méthodes y compris dans l'inférence causale sont utilisées à la marge comme la forêt aléatoire pour construire un score de propension ou voient le jour comme l'apprentissage ciblé (targeted learning). De ce fait, l'apprentissage automatique n'est pas considéré dans cette partie de la thèse.

2.1.2. Méthodes guidées par les données dans l'inférence causale

Les méthodes de sélection des variables guidées par les données qui semblent être les plus utilisées dans les études observationnelles pour la construction de modèles explicatifs sont les suivantes : d'une part l'analyse univariée entre le critère de jugement et les covariables et d'autre part le « change-in-estimate » (Talbot D et al 2019, Liao H et al 2010, Walter S et al 2009). (i) La première inclut les variables conditionnellement à un seuil de *p*-valeur qui peut varier entre 0,05 et 0,25 selon les cas ; (ii) La deuxième inclut les variables conditionnellement à un seuil de changement de l'estimation du coefficient de l'exposition d'intérêt induit par le retrait d'une covariable d'ajustement. Par exemple, soit le coefficient de mon exposition à 2,5, si je retire la variable 2 du modèle et que son retrait entraîne une modification de 0,25 du coefficient de mon exposition, alors je garde cette variable. Contrairement à la première méthode, cette dernière méthode est spécifiquement utilisée pour construire des modèles pour l'inférence causale. En effet, elle permet de détecter les variables qui sont liées à la fois à l'exposition et au critère de jugement.

D'autres méthodes de sélection des variables guidées par les données ont été développées pour construire spécifiquement ce type de modèles

La méthode de sélection appelée « purposeful » (Bursac Z et al 2008), est une méthode hybride entre 1- l'analyse univariée (par exemple avec un seuil de *p*-valeur <0,25), suivi de 2- pas-à-pas descendant qui permettent l'exclusion de variables (par exemple avec un seuil de *p*-valeur <0,10) qui ont aussi un « change-in-estimate » trop petit (par exemple <0,15) pour finalement 3- ré-inclure les variables retirées dans l'étape 1 selon les critères de l'étape 2.

La méthode de sélection appelée « augmented backward elimination » ou ABE (pas-à-pas descendant augmenté) est elle aussi une méthode hybride qui mêle le pas-à-pas descendant et le change-in-estimate [Dunkler D et al 2014]. Il est précisé que l'utilisation de ces méthodes doit être couplée avec les connaissances initiales. Dans l'article qui présente la méthode ABE, les auteurs définissent le jeu initial de variables avec le « disjunctive criterion », méthode basée sur la représentation des connaissances développée plus loin [VanderWeele TJ et al 2011]. Ce serait donc une méthode hybride guidée par les données et les connaissances.

Certaines méthodes guidées par les données reposent sur la construction automatique de modèles prédictifs pour répondre à des questions de causalité. Dans l'inférence causale, le score de propension est une méthode d'ajustement régulièrement utilisée. Le score de propension consiste à créer un modèle prédictif de l'exposition et de se servir de la probabilité

prédite pour ajuster un modèle, ou pour pondérer ou apparier les individus. Il peut être construit de manière automatique, grâce aux méthodes citées plus haut, mais aussi des méthodes provenant du domaine de l'apprentissage automatique comme les arbres de régression ou les forêts aléatoires. Une méthode a été spécifiquement créée pour les bases de données médico-administratives (base de données comptables) : le « high dimensional propensity score » ou score de propension à haute dimension [Schneeweis S et al 2009] qui évalue la *prévalence des variables*, leur *récurrence* chez un même patient, les priorise avec une méthode automatique pour finir par choisir les plus importantes. Celles-ci, serviront à construire le score de propension qui permettra par exemple d'ajuster directement ou d'apparier les deux groupes à comparer.

L'utilisation de méthodes d'estimation doublement robustes (*doubly robust estimate*) [Funk MJ et al 2011] repose sur la construction de plusieurs modèles prédictifs. Les modèles prédictifs de cette méthode peuvent être construits grâce à différentes méthodes guidées par les données comme l'adaptive group-lasso [Koch B et al 2018].

Les méthodes guidées par les données peuvent aussi être utilisées dans un but de réduire drastiquement le nombre de variables candidates dans des cas de haute dimension. Ces méthodes sont appelées « screening-based » dans la revue de Desboulets L. On retrouve parmi les plus connues l'analyse en composante principale. Elles ne seront pas plus détaillées ici mais se développent déjà dans un but de sélectionner les facteurs de confusion [Fan J et al 2008, Lee S et al 2019].

Interaction

Pour trouver les interactions, les méthodes guidées par les données pourront dans un premier temps rechercher les interactions significatives (p-valeur<X), puis dans un second temps, comparer l'apport de ce terme d'interaction dans le modèle en comparant le modèle sans et le modèle avec terme d'interaction (comparaison de modèles emboîtés).

Recodage

Avant de recoder une variable, un test de linéarité ou une comparaison de modèles pourra être effectué pour décider du recodage en variable quantitative ou variable qualitative (variable qualitative ordonnée à k modalités recodée en k-1 variables binaires). Un pas-à-pas descendant pourrait être utilisé pour choisir le nombre de k-nots lorsqu'on utilise des splines. Il est possible de faire des régressions en morceaux et laisser un algorithme décider des points d'inflexion. Dans le cas où il faudrait binariser la variable, un cut-off optimal pourra, par exemple, être trouvé via la maximisation de l'indice de Youden (sensibilité + spécificité -1).

2.1.3. Limites des méthodes guidées par les données

Limites générales des méthodes guidées par les données

Depuis plusieurs années déjà les méthodes basées sur les tests sont critiquées. En effet, elles mènent à (Thompson B et al 1989, Whittingham MJ et al 2006, Flom PL 2007) : (i) des erreurs standard (écart type moyen) trop bas, (ii) des *p*-valeurs faussement basses, (iii) des intervalles de confiance trop étroits notamment lors d'utilisation de méthodes classiques de calcul tel que le score de Wald et le rapport de vraisemblance, (iv) augmentent les problèmes de colinéarité, (v) des fluctuations d'échantillonnage non normales. Pour ce dernier point, on pourrait même parler de fluctuations des coefficients chaotiques puisque tantôt les coefficients ont une valeur tantôt la variable est retirée ce qui revient à lui attribuer la valeur 0 (pour une régression linéaire) ou 1 (régression logistique). Des auteurs comme Peter L. Flom proposent d'utiliser d'autres méthodes comme le LASSO ou la « leastangle regression » (LARS) qui sont supposées être de meilleures alternatives. Cependant, les méthodes basées sur les tests et basées sur la pénalisation ont des problèmes de validité identique lorsque les méthodes basées sur la pénalisation choisissent la pénalisation avec une méthode guidée par les données. En plus des problèmes de validité cités ci-dessus, elles partagent toutes deux les mêmes problèmes que les statistiques uniquement basées sur les *p*-valeurs (même si un AIC ou un BIC sont utilisés à la place) : un risque de fausse découverte (false discovery rate) augmenté par la multiplicité des tests (Colquhoun D 2014), l'acceptation inappropriée de l'hypothèse nulle lorsqu'un effet est non significatif (Benjamin DJ et al 2018), le phénomène de *p*-hacking aussi appelé data-dredging qui consiste à utiliser différentes méthodes et sous-jeux de données pour obtenir un $p<0,05$, le *p-harking* qui consiste à inventer une explication douteuse à un résultat statistiquement significatif.

Pour finir, la sélection des variables s'appuie sur la qualité des données dont on dispose (erreur de mesure fortement probable ou bruit supérieur au signal) et sur la chronologie des évènements (il est déconseillé de prédire le futur avec le futur). Une machine n'a pas conscience de ces deux points et peut donc faire des erreurs grossières dans ce sens.

Concernant le recodage ou le choix de seuil (*cut-off*) pour des variables quantitatives il a été démontré la grande variabilité des résultats lors de l'utilisation de méthodes guidées par les données (Bhandari PM et al 2021).

Cas particulier des tests visuels. Il est possible de tester visuellement si un effet est oui ou non significatif ou si une variable est plutôt linéaire ou non. Il suffit de faire un graphique de type histogramme ou nuage de points pour s'en apercevoir. Ce genre de tests passe

complètement inaperçu et pourtant est aussi voire plus délétère que le genre basé sur la p-valeur. Ils peuvent s'imposer dans le processus d'appropriation et de contrôle de cohérence des données ou être réalisés consciemment. Évidemment, les tests visuels rentrent dans la catégorie des approches guidées par les données.

Limites des méthodes guidées par les données, *dans le cadre de* l'inférence causale

Dans le sous chapitre qui suit sur les méthodes guidées par la connaissance, une dernière limite majeure des méthodes guidées par les données sera expliquée. Cette limite est présente lorsque l'on veut répondre à une question de recherche causale UNIQUEMENT. Les méthodes guidées par les données détectent des associations statistiques et non le sens du lien causal. Pour une méthode guidée par les données toutes les situations suivantes sont équivalentes : A cause B, B cause A, A et B partagent un ancêtre commun ou A et B ont une conséquence commune constante dans l'échantillon. Puisqu'elle ne fait pas cette différence, elle ne fait pas non plus la différence entre variables de confusion, médiation ou collision (ces concepts seront détaillés plus bas). De ce fait, elle ne fait pas la différence entre variable qui corrige le biais et qui crée du biais, ou autrement dit, elle ne fait pas la différence entre variable indispensable à sélectionner et variable prohibée pour la sélection.

2.2. Méthodes guidées par les connaissances

L'expression « sélection guidée par les connaissances » n'est pas un terme consacré. Aussi, il est important de définir ce que cela signifie. Le terme « connaissances » concerne les connaissances au sens large : connaissance théorique, pratique médicale, type d'étude, définition des variables, leur qualité, la quantité de données manquantes. Elle exclut la connaissance que le biostatisticien peut acquérir en réalisant certaines des figures descriptives comme un histogramme ou encore un graphique de corrélation à partir du jeu de données ou en exécutant des tests statistiques. En effet, utiliser cette connaissance là pour sélectionner les variables, correspond à prendre une décision dépendante des données, autrement dit, guidée par les données. Quand on parle de guider par les connaissances, ce sont donc les connaissances *a priori du début de l'analyse de la base de données* exception faite du pourcentage de données manquantes par variable qui pourra conditionner certains choix. La sélection guidée par les connaissances *a priori* englobe la sélection des variables à partir des connaissances disparates et peu ordonnées et la sélection à partir de représentations des connaissances.

En épidémiologie, les représentations de connaissances les plus utilisées sont le graphe orienté acyclique causal (DAG) ou le diagramme causal (CD).

2.2.1. Graphes orientés acycliques (DAGs) causaux

Définition

Dans les graphes orientés acycliques (DAGs) causaux, les variables sont représentées par des nœuds et les relations causales entre variables par des flèches (ou arêtes dirigées). Le DAG est **dirigé** car toutes les arêtes indiquent une direction. Si *x*, la variable d'exposition d'intérêt, cause *y*, la variable critère de jugement, alors ces variables seront représentées par deux nœuds et une flèche partant de *x* dirigée vers *y*. On dit que *x* est l'ancêtre (ici le parent direct) de *y* et *x* son descendant (ici l'enfant). En médecine, il est rare d'avoir un seul ancêtre. Par exemple, la taille à l'âge adulte est dépendante à la fois de la génétique et de notre environnement. Le DAG est **acyclique** car il n'existe aucune boucle fermée. Une variable *x* ne peut pas, par l'intermédiaire d'autres variables, se causer elle-même.

Les trois types de variables

A partir du moment où une troisième variable entre en jeu, nous pouvons définir de manière grossière trois types de variables résultants des relations causales :

- S'il existe une variable qui cause à la fois l'exposition et le critère de jugement, alors c'est une variable de **confusion** (figure 3) ;
- S'il existe une variable causée par l'exposition et que cette variable cause le critère de jugement, alors c'est une variable de **médiation** (figure 3) ;
- Un chemin causal est dit *bloqué* s'il existe une ou plusieurs variables de **collision** sinon il est dit non bloqué (ouvert). Schématiquement, une variable de collision est le « lieu » d'une collision entre deux relations causales qui pointent sur la même variable. Dans les figure 3 et 4 la variable de collision (Collider) bloque un chemin causal entre l'exposition et le critère de jugement.

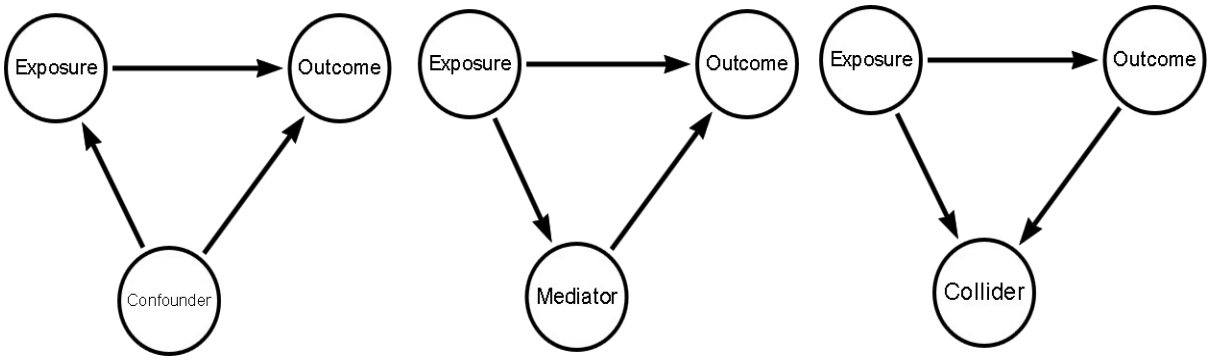


Figure 3. De gauche à droite, variable de confusion, variable de médiation et variable de collision.

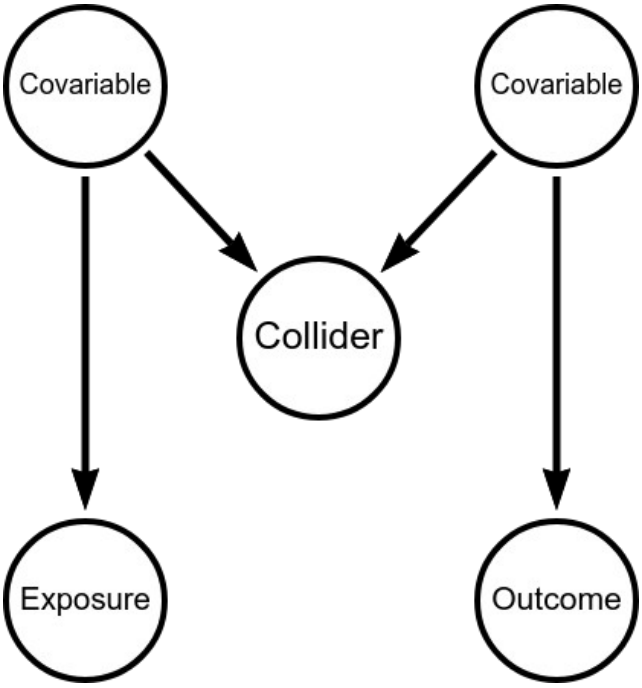


Figure 4. M-biais, situation avec une variable de collision formant un M.

Impact théoriques des ajustements et exemples

Ajustement sur une variable de confusion. Ne pas ajuster sur une variable de confusion entraînera un biais d'estimation de l'effet de la variable d'exposition d'intérêt sur le critère de jugement (biais de confusion). Un exemple courant de variable de confusion dans la pratique clinique est une variable d'indication d'un traitement :

Soit le critère de jugement 'décès dans la semaine suivant l'admission', la variable 'traitement' (A ou B) et la variable 'hémorragie' (aucune-limité-moyenne-massive). L'hémorragie massive du patient indique dans la majorité des cas le traitement A et augmente le risque de décès dans la semaine suivant l'admission. Considérons la question de l'évaluation de l'effet causal du traitement A par rapport au traitement B sur le risque de décès dans la semaine suivant l'admission. Sans ajustement sur le degré d'hémorragie du patient on pourrait croire à tort que le traitement A est moins efficace que le traitement B.

L'effet de la randomisation visualisé avec des DAGs : Précédemment, nous avons vu que la randomisation était la clé du contrôle de la confusion. Ce contrôle peut être représenté via les DAG : une situation sans randomisation figure 5 (étude observationnelle) et une situation avec randomisation figure 6 (essai clinique contrôlé randomisé). Dans la figure 5, on note la présence de plusieurs variables de confusion qui permettent à la fois de choisir le fertilisant et qui impactent le rendement, alors que dans la figure 6, elles ne permettent plus de choisir

le fertilisant. En effet, le choix du fertilisant ne dépend plus que du tirage aléatoire. Le biais de confusion est corrigé.

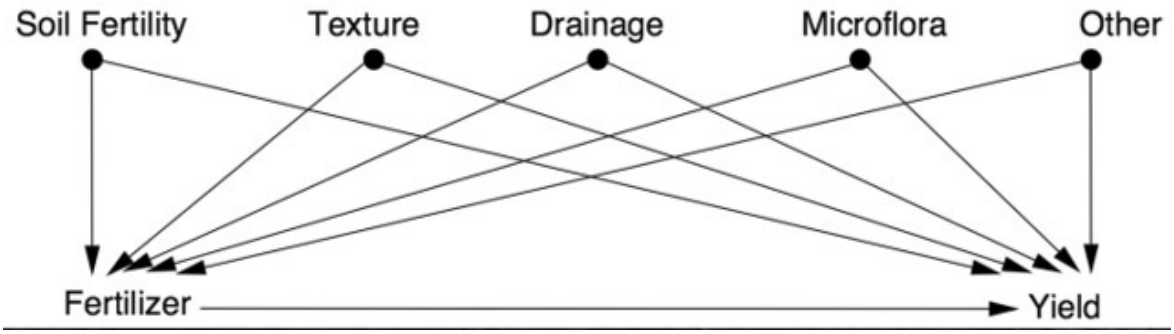


Figure 5. Une expérience mal contrôlée (The book of why de J Pearl, D Mackenzie 2019).

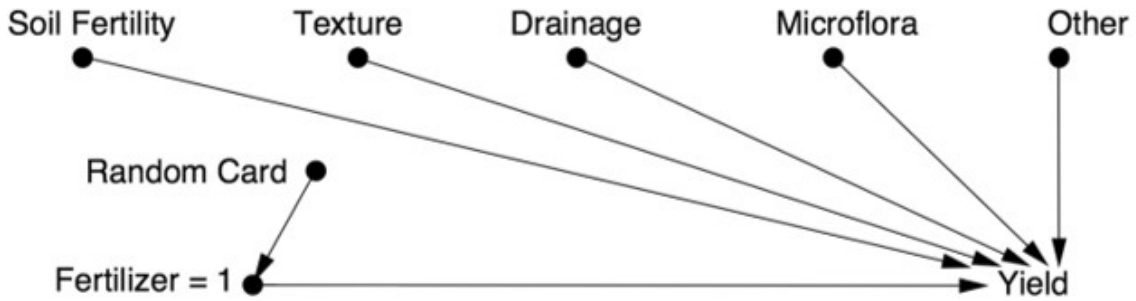


Figure 6. Le monde simulé par un essai contrôlé randomisé (The book of why J Pearl & D Mackenzie 2019).

Ajustement sur une variable de médiation. Ajuster sur une variable de médiation entraînera aussi un biais d'estimation :

Soit le critère de jugement infarctus du myocarde et le groupe de facteurs de risque suivant : obésité, sédentarité, régime alimentaire et diabète. L'obésité est en partie causée par la sédentarité et le régime alimentaire. Considérons la question de l'évaluation de l'effet causal de la sédentarité sur le risque d'infarctus, dans l'idée de favoriser les politiques de réduction de la sédentarité si cet effet s'avérait important. L'ajustement sur l'obésité, facteur de médiation, conduira à une sous-estimation de l'effet de la sédentarité sur l'infarctus du myocarde car seul l'effet direct de la sédentarité sera calculé par le modèle (Figure 7).

Cependant, l'ajustement sur l'obésité peut être intéressant pour étudier l'effet médié par celle-ci. L'effet direct de la sédentarité est obtenu après ajustement sur l'obésité, l'effet total sans ajustement sur l'obésité et l'effet indirect médié par l'obésité en soustrayant total et direct. Il est aussi possible d'obtenir l'effet indirect par le produit des coefficient $Beta_1$ et $Beta_2$ et

obtenir l'effet total par addition de l'effet direct et l'effet indirect (VanderWeele TJ et al 2009).

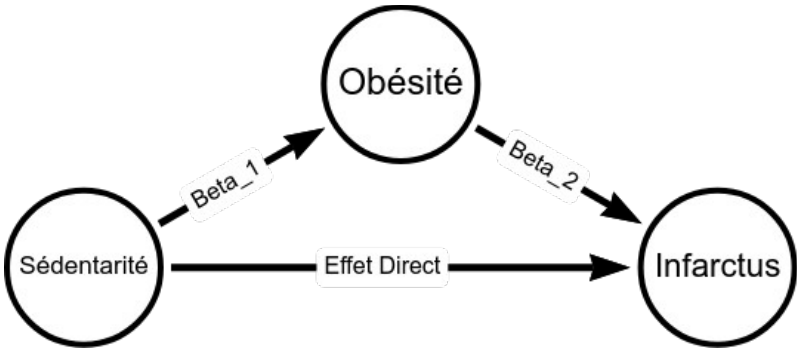


Figure 7. Effet direct et indirect dans une situation de médiation.

Ajustement sur une variable de collision. Ajuster sur une variable de collision entraînera un biais d'estimation de l'effet de la variable d'exposition d'intérêt sur le critère de jugement (Berkson J 2014, Sackett DL 1979). Un des exemples classiques est celui de Berkson aussi discuté par Sackett appelé par ce dernier « admission rate biais » :
Considérons la variable 'maladie 1', la variable 'maladie 2' indépendantes l'une de l'autre et la variable de collision 'hospitalisé' causée par la variable maladie 1 et la variable maladie 2. Dans le cas où nous ajustons sur la variable 'hospitalisé', la relation entre les deux variables maladies est créée. C'est à dire que d'un point de vue statistique, il existe un lien non nul entre ces deux maladies. Cette erreur peut mener à conclure que la maladie 1 protège de la maladie 2.

Sur la figure 4 représentant le M-biais (Greenland S et al 1999), la variable de collision n'est pas entre la variable d'intérêt et le critère de jugement mais entre deux prédicteurs indépendants de la variable d'intérêt d'une part et du critère de jugement de l'autre. Dans ce cas, si la variable de collision est une variable d'ajustement, cela entraîne toujours un biais. Il peut, en revanche, être corrigé en ajustant sur un des sommets (nœuds).

Voici un dernier exemple sur un biais de sélection retrouvé en chirurgie. Soient deux variables 'taille de tumeur bénigne' et 'symptomatologie' et une constante 'patient opéré de de la tumeur' (oui). Normalement, plus la taille est importante plus la symptomatologie est bruyante. Les patients opérés sont ceux avec une symptomatologie bruyante ou avec une taille importante de tumeur. Cela entraîne que dans le groupe des opérés, quand on a une petite tumeur, on a une forte symptomatologie (Figure 8)

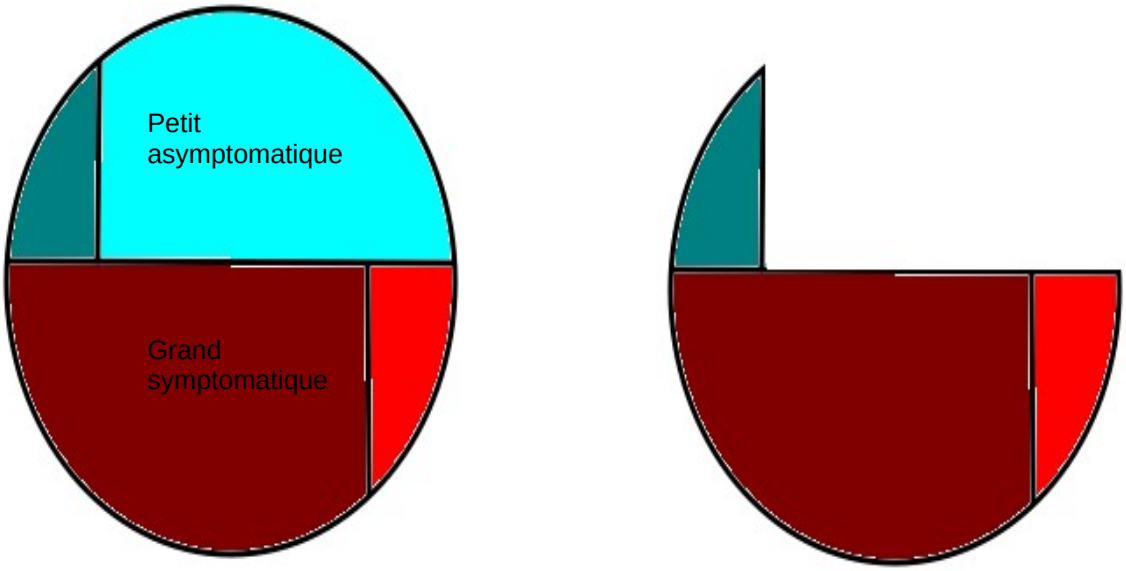


Figure 8. Représentation du biais de collision. Première figure représentation de la part de patient symptomatique parmi les patients avec une tumeur de grande taille et la part des patients asymptomatiques parmi les les patients avec une tumeur de petite taille. Deuxième figure, représentation des parts après sélection des patients opérés : patients symptomatiques ou grande tumeur.

Ajuster sur une variable de confusion peut entraîner un biais. Si nous reprenons l'exemple du M-biais et que la variable de collision est aussi une variable de confusion, il faut ajuster sur celle-ci. Comme vu dans la partie ajustement sur une variable de collision, cet ajustement va créer un lien entre les deux sommets du M et donc va faire de ces sommets des variables de confusion. Il faudra ajuster sur un des deux sommets en plus de la variable de confusion/collision pour corriger les biais (Figure 9).

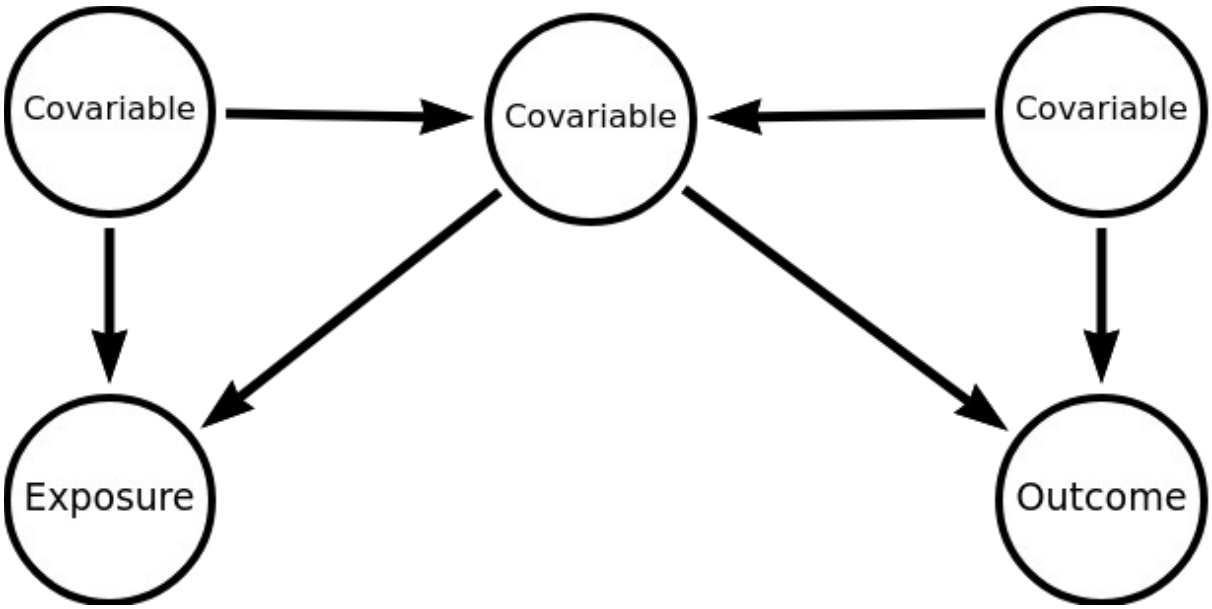


Figure 9. Représentation d'une variable de collision qui est aussi une variable de confusion.

Une variable de confusion, qui est aussi une variable de médiation, est un véritable problème qui peut être géré en évaluant les pour et les contre de l'ajustement en fonction de l'impact anticipé mais aussi en réalisant des analyses de sensibilité. Les analyses de sensibilité sont cruciales en causalité car elles permettent d'estimer l'effet causal sous plusieurs scénarios différents. Si et seulement si les résultats convergent alors la conclusion est unanime sinon elle sera beaucoup plus nuancée.

La dernière limite des méthodes guidées par les données dans l'inférence causale vient d'être démontrée avec une représentation des connaissances simpliste. En effet, les méthodes guidées par les données ne distinguant pas confusion, collision et médiation, elles peuvent choisir à tort ces deux derniers types de variables. Après les DAG causaux, nous allons passer à un autre type de représentation très proche, appelé diagrammes causaux.

2.2.2. Les diagrammes causaux

Le diagramme causal est plus souple qu'un DAG en terme de représentation des liens entre variables. Deux nouveaux liens sont utilisés : associations dues à un ancêtre commun (arêtes bidirectionnelles) ou associations dues à une conséquence commune constante dans le jeu de données (arêtes non dirigées) (Greenland S et al 1999).

Relations dues à un ancêtre commun

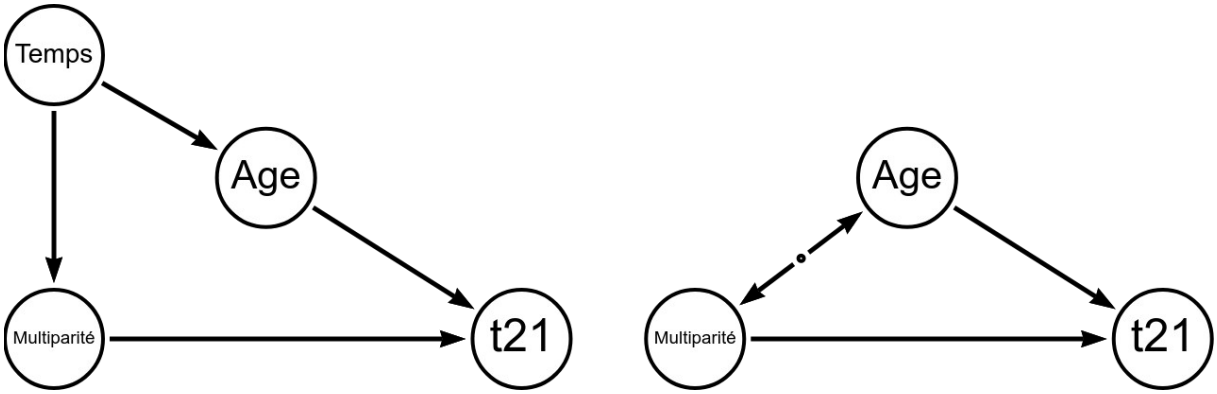


Figure 10. A gauche la variable temps est un ancêtre commun de la multiparité et de l'âge; à droite la variable temps est supprimée et un arc bidirectionnel est crée.

Lorsque deux variables sont associées, cela peut venir du fait qu'elles partagent un ancêtre commun. C'est très courant en médecine où nombre de maladies ou comportements de santé ont un ou plusieurs déterminants communs. Les arcs ne sont plus unidirectionnels mais bidirectionnels pour signifier la présence d'un ancêtre commun non représenté par un nœud (figure 10). La flèche bidirectionnelle permet de prendre en compte des variables latentes, ou variables non observées, ou facteurs de confusion non mesurés sans avoir besoin

de les représenter par un nœud (Greenland S et al 1999). Ce sont des variables qui, pour différentes raisons, n'ont pu être recueillies dans le jeu de données.

Relations dues à une conséquence commune constante

Il arrive que deux variables soient associées sans pour autant partager d'ancêtre commun. Ces relations (ou liens) peuvent être représentées par des arcs non dirigés en pointillés. Ces relations peuvent découler de sélection de populations particulières. Soit l'exemple de la variable de collision (*m*) de la figure du M biais (Figure 4). Dans une population sélectionnée sur un *m* positif (exclusion *m* négatif), il existe une association non causale entre A et B que l'on pourra représenter par un arc non dirigé (éventuellement en pointillés) (Figure 11).

Ces deux nouvelles relations permettent une schéma plus synthétique avec moins de nœuds et de flèches (i.e., 3 nœuds et 2 flèches dans un DAG peuvent devenir 2 nœuds et une flèche dans un diagramme causal).

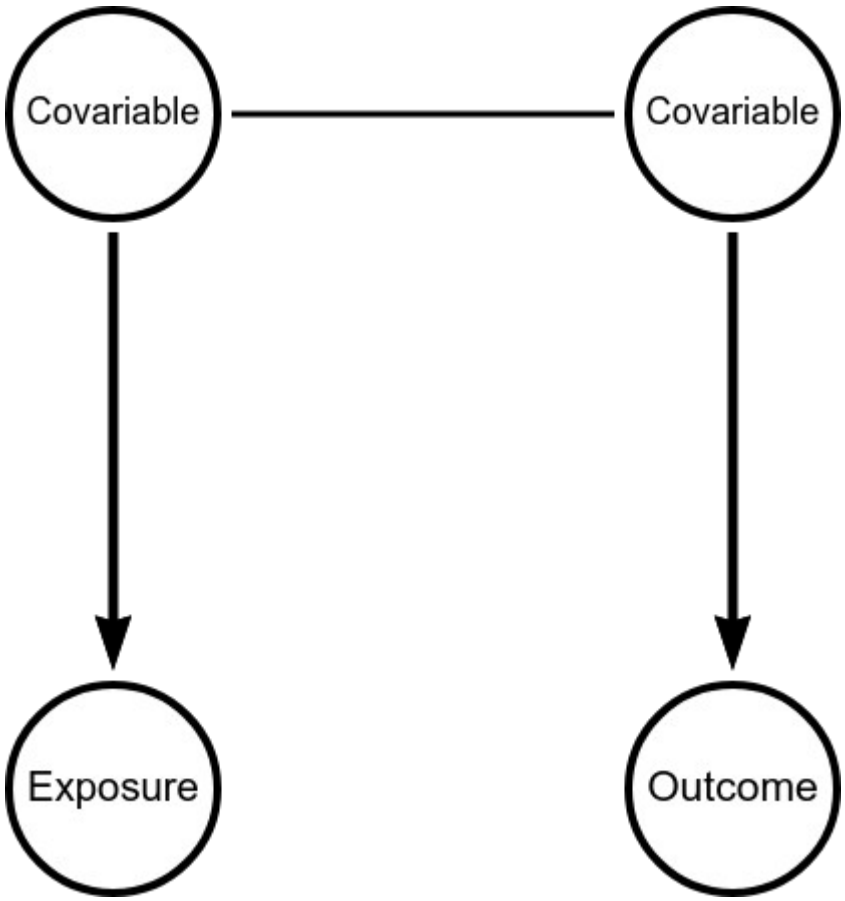


Figure 11. Ajustement sur une variable de collision entraînant la création d'un arc non dirigé entre deux covariables.

Les biais qui ne répondent pas à la définition de biais de confusion

Comme présenté dans la partie biais de sélection et de mesure des études observationnelles, il arrive que les biais différentiels découlent directement de l'exposition ou même du critère de jugement. Par exemple, dans une étude cas-témoins, l'enquête sur l'exposition passée chez un cas peut parfois être plus fine et longue que chez un témoin menant à un biais de mesure. Dans cette situation l'« enquête méticuleuse » directement causée par le fait d'être malade cause l'exposition, ce qui ne respecte pas la définition de variable de confusion mais qui correspond à un biais de mesure. Inversement, chez un exposé on aura tendance à plus surveiller sa santé et donc détecter un critère de jugement de manière plus précoce. Ici, c'est bien le fait d'être exposé qui conduit à le suivre et le suivi amène à détecter de manière plus précoce la maladie. Le suivi se place en variable de médiation, pourtant, dans cette situation il biaise l'effet de l'exposition à proprement parler. C'est un biais de suivi qui entre dans la catégorie des biais de sélection.

Biais impossibles à corriger

Dans certaines circonstances, les biais ne peuvent pas être corrigés et ne sont pas forcément visibles car les variables deviennent des constantes lorsque le recrutement des patients l'impose. Reprenons deux des exemples précédents :

- Concernant le lien causal entre sédentarité et infarctus du myocarde. Si les patients atteints d'obésité sont exclus, alors l'impact sur l'estimation du lien causal entre la sédentarité et l'infarctus sera comparable à l'ajustement sur la variable de médiation obésité ;
- Concernant l'absence de lien causal entre la maladie 1 et la maladie 2. Si les patients non hospitalisés sont exclus, alors l'impact sur l'estimation du lien causal entre maladie 1 et 2 sera comparable à l'ajustement sur la variable de collision hospitalisé ;
- Pour comparer deux traitements, quel que soit le patient, il faut qu'il puisse bénéficier soit de l'un soit de l'autre. Si ce n'est pas le cas la contre indication dite absolue peut rendre impossible la comparaison des traitements.

2.2.3. Méthodes de sélection des variables d'ajustement basée sur les DAGS et/ou les diagrammes causaux

Une fois le DAG créé, diverses méthodes existent pour sélectionner les variables à partir de cette représentation. Certaines peuvent être résolues manuellement, d'autres automatiquement. Voici les plus connues :

- L'approche qui respecte la définition de facteur de confusion en ajustant sur les variables qui causent à la fois l'exposition d'intérêt et le critère de jugement (« common cause ») (Glymour MM et al 2008);

- L'approche « canonique » qui inclut toutes les variables ancêtres de l'exposition ou du critère de jugement ou des deux, qui ne sont pas des variables descendantes de l'exposition (i.e, causée par l'exposition). (Cette méthode est nommée « canonical » dans le package R Dagitty [Textor J et al 2017).

- L'approche du « back-door criterion » par Judea Pearl 1993: Soit une variable d'exposition d'intérêt X et le critère de jugement Y. Un jeu de variables Z satisfait le « back-door criterion » s'il n'y a pas de descendants de X dans Z et si Z bloque tous les chemins entre X et Y qui contiennent une flèche dirigée vers X. Autrement dit la première assertion évite les problèmes de médiation et la seconde corrige les biais de confusion et évite d'ouvrir un chemin en ajustant sur un facteur de collision. Cette approche permet de sélectionner le jeu ou les jeux minimaux de variables ;

- L'approche du « disjunctive criterion » (VanderWeele TJ et al 2011) qui consiste à inclure les variables qui causent l'exposition d'intérêt ou le critère de jugement ou les deux. Lorsque le diagramme causal est incomplet, elle serait plus adaptée que les techniques précédentes (VanderWeele TJ 2019). Une variante de cette approche propose d'écarter les variables dites instrumentales et de conserver les bons proxys (*modified disjunctive criteria* (VanderWeele TJ 2019).

2.2.4. Conclusion

Pour se rapprocher du vrai effet causal, il faut donc supprimer les biais de confusion tout en évitant les biais dûs à la sélection des variables de médiation ou de collision. Les DAG sont une des solutions qui permettent de se prémunir contre les erreurs de sélection. Inclure une variable indépendante du critère de jugement principal (correction par excès) diminuera la précision de l'estimation [Greenland S 2007], ce qui est beaucoup moins préjudiciable que l'oubli d'une variable de confusion.

Les règles de sélection basées sur les connaissances sont nombreuses (*disjunctive criterion*, *backdoor,* etc). Étant donné que la correction par excès est peu coûteuse, les méthodes telles que le *disjunctive criterion* semblent représenter une bonne option. De plus, c'est une méthode qui a l'avantage d'être claire et facilement explicable, contrairement au *backdoor criterion* qui propose un jeu minimal ou plusieurs à sélectionner. Elle est aussi plus facile d'utilisation lorsque l'on dispose uniquement d'une représentation visuelle. Cependant,

utiliser plus de variables c'est prendre le risque d'intégrer des erreurs dues à la qualité de la mesure, mais aussi de diminuer la puissance à cause des données manquantes.

Limites des méthodes basées sur les connaissances

Les DAG sont construits à partir des connaissances. Les DAG partagent donc les limites des connaissances qui peuvent parfois être purement hypothétiques et évolutives.

Limites de construction. La construction des DAG reste une activité assez libre et peu formelle mise à part un nœud = une variable, une flèche = une relation causale. Cela a un avantage incontestable lorsque l'on apprivoise pour la première fois cette méthode. Certains auteurs ont proposé des tutoriels afin de faire attention aux pièges lors de la construction ou afin d'obtenir des constructions plus systématiques (Suzuki E et al 2020).

Limites de représentations. Du fait de cette libre construction, la représentation n'est pas expressive. L'information contenue dans les DAG est très basique et peut parfois être largement suffisante pour la tâche de sélection des variables. Néanmoins, il est important de noter que les connaissances étant plus complexes, il arrive que la représentation soit trop superficielle (du moins pour les moins aguerris). On pourrait donc imaginer enrichir la représentation des DAG avec par exemple la notion de pourcentage de données manquantes par variable, ou encore la forme du lien (linéaire ou exponentielle) entre deux variables.

La pratique des DAG. Aujourd'hui les DAG sont rarement utilisés et jamais réutilisés (même de manière parcellaire) et rarement réutilisables.

Limites pour la sélection des variables. Dans certains cas particuliers, la représentation avec un DAG classique peut conduire à une mauvaise sélection des variables (e.g., biais de suivi).

Synthèse du chapitre

Il existe de nombreuses méthodes pour la sélection des variables dans l'inférence causale. Parmi elles, les méthodes automatiques guidées par les données semblent par essence ne pas être adaptées à l'inférence causale. Néanmoins, elles sont encore utilisées, mais peu dans cinq des journaux ayant le plus gros facteur d'impact dans la catégorie médecine. (Pressat-Laffouilhère T et al 2021). Les DAGs causaux et les diagrammes causaux sont des représentations interchangeables simplistes des connaissances qui permettent la sélection du jeu minimal de variables pour dé-biaiser le vrai effet causal. Ils sont la référence pour sélectionner les variables de confusion et éviter les variables de médiation et de collision mais les méthodes de construction restent insuffisamment systématiques, la représentation n'est pas

expressive (e.g., logique de description absente de la représentation), insuffisamment formelle, les DAG des études sont peu partagés et non réutilisés.

Hypothèse de travail

L'utilisation d'une ontologie avec des règles d'inférences pourrait, en tant que système expert, palier les quelques faiblesses des DAG et guider les biostatisticiens, et les cliniciens tout au long du processus de sélection des variables. En effet, les ontologies permettent une représentation des connaissances avec une haute expressivité. Elles peuvent être utilisées comme un modèle de données, dans notre cas, un canevas pour la construction de DAG. Ce modèle est décomposé en classes, relations (*object properties*) et *data properties*. Les ontologies rendent explicite une partie des connaissances qui est jusqu'alors implicite ; permettent de partager avec les autres chercheurs une représentation aux standards communs ; et la réutilisation de certaines représentations ou sous partie de représentations. De plus, les ontologies font partie de la famille des intelligence artificielles symboliques qui, contrairement aux intelligence artificielles basées sur l'apprentissage profond, fournissent nativement des explications sur les résultats (évitent le phénomène de « boîte noire », ce qui est indispensable en médecine). En effet, les systèmes basés sur la logique sont capables de fournir le raisonnement qui a conduit à une conclusion donnée. Grâce à une combinaison de règles logiques et d'axiomes, elles permettraient d'aider à la compréhension et l'analyse des diagrammes causaux.

Chapitre 3 : Ontologie et sélection de variables dans l'inférence causale dans la recherche biomédicale : état de l'art

Avant la création de l'outil ontologique, (i) nous définirons d'abord ce qu'est une ontologie, (ii) quels sont ses composants basiques, (iii) les composants plus avancés en partant du langage rdf comme base secondairement enrichi par OWL dont les limites de raisonnement sont compensées par l'utilisation de SWRL, (iv) comment utiliser une ontologie avec des règles ou des requêtes, (v) les différents types d'ontologies seront présentés afin de situer celle qui sera construite, (vi) quelques une des méthodes de constructions seront détaillées afin de comprendre l'intérêt de celle qui sera choisie, (vii) quels outils permettent de créer une ontologie et pour finir (viii) quelles sont les différentes façon d'évaluer une ontologie. Ensuite, un état de l'art concernant le sujet de la ***sélection des variables pour la création de modèles multivariés*** *sous forme ontologique sera réalisée.*

1. Ontologie : généralités

1.1. Définition

Chaque individu a une conception du monde ou un modèle ontologique du monde ou vision du monde (i.e., comment il est organisé). Nous nous appuyons tous sur celui-ci pour raisonner. Par exemple, lorsque nous parlons d'organes du corps humain, nous savons la liste des objets qui sont et qui ne sont pas des organes. Afin d'améliorer les échanges entre machines et individus, il faudrait fournir aux machines une représentation du monde qui pour nous est très largement acquise, voire implicite, contrairement aux machines.

L'ontologie 'informatique' ne doit pas être confondue avec l'ontologie philosophique (branche de la métaphysique qui étudie la signification de l'être). L'ontologie informatique a été définie de bien des manières telles que : spécification **explicite** d'une **conceptualisation partagée** pour un **domaine de connaissance** [Gruber et al 1993] :

(i) **Explicite** signifie que les concepts utilisés ont une définition explicite (e.g., les chaises ont un dossier et quatre pieds) ;

(ii) **Domaine de connaissance** veut dire que cela s'applique pour un domaine en particulier comme la médecine ou l'immobilier ou un domaine plus général, comme la causalité ;

(iii) **Conceptualisation** fait référence à une conception qu'on peut se faire du monde et notamment d'un domaine de connaissance. C'est un modèle dans lequel s'articulent les concepts propres à ce domaine de connaissance.

(iv) **Partagée** signifie que la conceptualisation de la connaissance est consensuelle au sein d'un groupe d'individus et va permettre des échanges compréhensibles entre individus.

Cette définition peut être complétée par celle de Borst [Borst 1997] qui mentionne une spécification formelle, ce qui signifie un langage compréhensible par une machine.

En pratique « *les ontologies représentent des théories sur différents objets, les propriétés de ces objets et les relations entre ces objets qui sont acceptables dans un domaine de connaissance spécifique* » [Chandrasekaran et al 1999].

Elles diffèrent des terminologies par la nature des relations entre objets. Dans une terminologie, on retrouve uniquement des relations dites hiérarchiques ou de subsomption. Par exemple, l'insuffisance cardiaque *est une* maladie du cœur. Dans une ontologie, des relations dites sémantiques viennent s'ajouter. Par exemple, la relation *est causée par* dans l'exemple : l'insuffisance cardiaque *est causée par* l'infarctus du myocarde.

1.2. Les constituants de base d'une ontologie

1.2.1. Concepts (ou objets ou classes)

Les concepts peuvent représenter un objet matériel (un coeur) ou immatériel (le temps). Ils sont choisis en fonction des objectifs de l'ontologie. Le concept est défini par des propriétés, des attributs, voire des contraintes (cardinalité) appelées **intentions**. L'intention se réfère à la définition abstraite et aux caractéristiques qui décrivent les membres d'une classe (concept) sans nécessairement spécifier les instances individuelles. C'est la signification ou la description conceptuelle de la classe. Prenons l'exemple de la classe "Fruit". Son intention pourrait inclure des caractéristiques telles que "organisme comestible produit par une plante", décrivant ainsi les propriétés communes de tous les fruits sans se référer à des exemples spécifiques. L'**extension** d'une classe dans une ontologie se réfère à l'ensemble concret d'instances individuelles qui appartiennent à cette classe. C'est la collection réelle d'objets ou d'entités spécifiques qui correspondent à la définition abstraite de la classe. Par exemple, si nous reprenons l'exemple de la classe "Fruit", l'extension de cette classe pourrait inclure des instances telles que "Pomme", "Banane", etc.

La différence entre instance et concept est subtile, mais la difficulté réside dans le fait de décider si un objet donné doit être représenté par un concept ou une instance d'un autre concept. Par exemple, le coeur pourrait être l'instance du concept organe, ou un concept subsumé par le concept organe.

Parallèle avec le triangle sémiotique :

Le triangle sémiotique est un concept introduit par le philosophe Charles Sanders Peirce pour décrire la relation entre le signe, le signifié et le référent. Il se compose de trois éléments principaux :

1. **Signe (*Sign*):** La représentation matérielle ou le symbole utilisé pour représenter quelque chose d'autre (Le fruit).
2. **Signifié (*Signified*):** La signification ou le concept associé au signe (organisme comestible produit par une plante).
3. **Référent (*Referent*):** L'objet ou la réalité du monde réel auquel le signe fait référence (Banane).

Les intentions/extensions d'une ontologie et le triangle sémiotique partagent l'idée de séparer la signification abstraite d'un concept de ses occurrences ou instances concrètes

1.2.2. Les propriétés

Il existe deux types de propriétés :

Les relations (*object properties*) relient les concepts (ou leurs instances) entre eux à la manière des arêtes dirigées qui relient les nœuds entre eux dans les DAG. Elles permettent aussi de relier des instances de concepts entre elles. Dans l'exemple précédent : « le **coeur** *est vascularisé* par des **coronaires** », *est vascularisé* est une relation qui relie **coeur** et **coronaires**. Comme les concepts, les relations sont définies par un terme.

Les attributs (*data type properties*) relient des concepts (ou instances) à un jeu de valeurs comme une date, un nombre ou un texte.

1.3. Du langage RDF à OWL jusqu'aux règles SWRL

Afin de développer plus amplement le contenu d'une ontologie, je commencerai par définir le standard RDF (Resource Description Framework) puis présenterai le langage OWL (Ontology Wen Language) pour finir sur le langage SWRL. Le contenu de cette partie est guidée par de nombreuses ressources disponibles sur le site https://www.w3.org/.

1.3.1. RDF et RDFschema

Le langage Ressource Description Framework est un standard développé par le World Wide Web consortium (W3C) afin de coder la connaissance des pages Web pour la rendre compréhensible pour la machine.

Dans l'intitulé de ce langage, (i) **Ressource** correspond à tout ce qui peut être identifié par une URI (« Uniform Ressource Identifier »). Une URI comprend les URL et permet d'identifier sur le web ce qui existe de notre monde avec un identifiant unique ; (ii)

Description correspond aux attributs, caractéristiques et relations entre ressources ; (iii) **Framework** correspond au modèle RDF dans sa globalité.

Le langage RDF est un modèle de triplet et de graphe : (i) La ressource (URI) est décrite par un triplet : (sujet ; prédicat ; objet). Par exemple, le **coeur** (sujet) *est vascularisé* (prédicat) par des **coronaires** (objet) ; (ii) Ce triplet peut être vu comme un graphe : dans l'exemple **coeur** et **coronaires** sont les sommets et *est vascularisé* est un arc ou une arête dirigée. Cette représentation peut se transformer en multigraphe en augmentant le nombre d'arcs entre deux sommets. Par exemple, les **coronaires** *se situent sur le* **coeur**. C'est un multigraphe orienté comme le DAG, car ses arêtes sont dirigées. Il est étiqueté, c'est à dire que les sommets et les arcs ont une étiquette sous la forme d'une URI ou d'un littéral (i.e., une ou plusieurs chaînes de caractères).

Le langage RDFschema permet de documenter le vocabulaire utilisé dans RDF avec une hierarchie de classe et une hiérarchie de propriétés.

Hiérarchie de classes : Grâce à la balise rdfs:class et rdfs:subClass, on va pouvoir dire que le **coeur** est une sous classe de la classe **organe**. Ceci permet d'inférer que toute instance de coeur est aussi un organe (propagation des types). De plus, toute classe est aussi sous classe d'elle-même (réflexivité de la subsomption). Si une classe **variable de confusion** est sous classe de **covariable** elle même sous classe de **variable**, alors **variable de confusion** est une sous classe de **variable** (transitivité de la subsomption).

Hiérarchie de propriété : grâce à la balise rdf:property, on peut typer *est vascularisé* comme une propriété (synonyme de relation et prédicat). La balise rdfs:subProperty permet de dire que *est une cause de* est une sous propriété de *relation causale*. Les inférences sont du même type que pour la hiérarchie de classes à savoir propagation des types, réflexivité de la subsomption et transitivité de la subsomption.

Les propriétés (relations) peuvent avoir une signature c'est à dire un domaine (*domain*) et un co-domaine (*range*). Le domaine (*domain)* correspond au concept de départ de la relation et le *range* correspond au concept vers lequel la relation pointe (concept d'arrivée). Par exemple, la relation *est vascularisé* devrait partir d'un concept correspondant à un **organe** vers un concept correspondant à des **vaisseaux sanguins** ce qui permet d'inférer que, lorsque la relation *est vascularisé* relie deux instances, celles-ci sont respectivement classées dans **organe** et **vaisseaux sanguins**.

Les propriétés et les classes peuvent avoir : (i) des étiquettes (*labels*), c'est-à-dire des noms compréhensibles par les humains plutôt que des codes ; (ii) des commentaires, c'est-à-dire des explications ou définitions en langage naturel.

Dans le paragraphe suivant sera présenté le langage OWL qui dépasse le langage RDFs en terme expressif au niveau sémantique et logique.

1.3.2. OWL

Le langage OWL (Ontology Web Langage) est un langage informatique basé sur la logique. Il étend la description du langage RDF afin d'obtenir de meilleures et nouvelles inférences. En 2004, OWL contenait d'abord trois sous langages que sont OWL Lite, OWL DL, OWL Full (classés par ordre croissant d'expressivité). OWL DL permet d'avoir un maximum d'expressivité tout en conservant des capacités de raisonnements complets et décidables. Puis OWL 2 a été créée en 2009 apportant trois nouveaux profils (OWL 2 EL, OWL 2 QL, OWL 2 RL) et une nouvelle expressivité (propriétés chaînées, propriétés asymétriques, réflexives, disjointes, etc). OWL 2 EL permet la résolution d'algorithmes en temps polynomial adapté aux très grandes ontologies. OWL 2 QL convient aux petites ontologies avec beaucoup d'individus où il est utile d'accéder aux données grâce à des requêtes relationnelles. OWL 2 RL est utile pour les ontologies légères avec un grand nombre d'individus où il est nécessaire de raisonner directement sur des données sous forme de triplets RDF (par exemple, Livre écritpar Auteur et Livre éditépar Editeur). Ci-dessous, OWL correspond à OWL 2 DL qui permet une très bonne expressivité et décidabilité.

Le langage OWL permet la définition de classes (concepts) via des opérateurs logiques en utilisant d'autres classes. La définition est : soit celle d'une classe **primitive** (*subClassof*) par une condition nécessaire ET NON suffisante ; soit une classe **définie** (*EquivalentClass*) par une condition nécessaire ET suffisante. De cette manière, il est aussi possible d'avancer que deux classes ou deux propriétés sont équivalentes, c'est à dire qu'elles rassemblent les mêmes instances (ou ressources dans le cadre du langage RDF). Il est aussi possible de spécifier que deux instances sont les mêmes (*sameAs*) ou différentes (*differentFrom*). Les opérateurs logiques regroupent : (i) Union (OU), (ii) Intersection (ET), (iii) Complément de (NOT) et (iv) Disjonction (aucun individu ne peut être l'instance de deux classes disjointes).

Le langage OWL permet de caractériser les relations (*object properties*) par des propriétés algébriques. Soit l'ensemble A représentant l'ensemble des instances ou des objets sur lesquels la relation est définie. Chaque élément de cet ensemble est une instance particulière, et la relation R spécifie comment ces instances sont liées les unes aux autres. Pour donner un

exemple concret, considérons un ensemble A qui représente les personnes, et une relation R qui indique si deux personnes sont des amis. Dans ce cas, A serait l'ensemble des individus (les instances), et la relation R déterminerait quels individus sont amis les uns avec les autres. Ainsi, chaque élément de A (chaque personne dans cet exemple) est une instance sur laquelle la relation R est définie, et les propriétés algébriques décrivent le comportement de cette relation par rapport à ces instances. Voici la définition, l'intérêt, et un exemple illustratif pour chaque propriété algébrique de relation :

Symétrique :

- *Définition :* Une relation est symétrique si, pour chaque paire d'éléments (a,b) dans la relation, l'élément (b,a) est également inclus.
- *Intérêt :* Elle exprime une relation de réciprocité entre les éléments.
- *Exemple :* La relation "être conjoints" est symétrique : si A est marié à B, alors B est marié à A.

Antisymétrique :

- *Définition :* Une relation est antisymétrique si, pour chaque paire d'éléments distincts (a,b) dans la relation, cela implique que (b,a) n'est pas inclus.
- *Intérêt :* Elle garantit qu'il n'y a pas de réciprocité stricte entre les éléments.
- *Exemple :* La relation "être parent de" est antisymétrique : si A est parent de B, alors B ne peut pas être parent de A.

Réflexive :

- *Définition :* Une relation est réflexive si chaque élément de l'ensemble est en relation avec lui-même.
- *Intérêt :* Elle indique que chaque instance est liée à elle-même dans la relation.
- *Exemple :* La relation "être identique à" est réflexive : chaque objet est identique à lui-même.

Irréflexive :

- *Définition :* Une relation est irréflexive si aucun élément n'est en relation avec lui-même.
- *Intérêt :* Elle exclut toute relation d'un élément avec lui-même.
- *Exemple :* La relation "être strictement plus grand que" est irréflexive : aucun nombre n'est strictement plus grand que lui-même.

Transitive :

- *Définition :* Une relation est transitive si, pour chaque paire d'éléments (a,b) et (b,c) dans la relation, cela implique que (a,c) est également inclus.
- *Intérêt :* Elle reflète une propagation logique des relations.
- *Exemple :* La relation "être parent de" est transitive : si A est parent de B, et B est parent de C, alors A est parent de C.

Fonctionnelle :

- *Définition :* Une relation est fonctionnelle si chaque élément de l'ensemble a au plus un élément associé dans la relation.
- *Intérêt :* Elle assure qu'un élément source a une seule cible.
- *Exemple :* La relation "avoir un numéro de sécurité sociale" est fonctionnelle : chaque personne a un unique numéro de sécurité sociale.

Inverse fonctionnelle :

- *Définition :* Une relation est inverse fonctionnelle si chaque élément de la cible a au plus un élément source associé dans la relation.
- *Intérêt :* Elle garantit qu'un élément cible a au plus un élément source.
- *Exemple :* La relation "être le patron de" est inverse fonctionnelle : chaque employé a au plus un patron.

Inverse :

- *Définition :* La relation inverse R−1 a les paires inverses par rapport à la relation R.
- *Intérêt :* Elle exprime la relation en sens inverse.
- *Exemple :* Si R représente la relation "être le père de", alors R−1 représente "être l'enfant de".

Disjointe :

- *Définition :* Deux relations sont disjointes si elles n'ont aucun élément en commun.
- *Intérêt :* Elle assure l'absence de chevauchement entre les deux relations.
- *Exemple :* Les relations "être parent de" et "être conjoints" sont disjointes : une personne ne peut pas être à la fois le parent et le conjoint d'une autre personne.

Les propriétés (relations ou attributs) utilisées pour la définition d'une classe peuvent être restreintes. Par exemple, on peut dire que le **coeur** est un organe avec exactement quatre cavités ou qu'un modèle statistique a au moins deux variables. Les différentes restrictions utilisables sont les suivantes : *only* (toutes les valeurs), *some* (certaines valeurs), *value* (une seule), *hasSelf* (soi même), maximum, minimum et nombre ou valeur exacte.

La définition d'un object property ou d'une classe avec une description logique utilisant des objets définis plus haut est appelée **axiome**. La définition de la classe variable de confusion serait : C'est une **Covariable** *qui Cause* à la fois le **Critère de jugement** et l'**Exposition**.

Avec le langage OWL, il n'est pas possible d'inférer la relation entre deux individus à partir de la relation qu'ils entretiennent avec un troisième. Le langage SWRL (Horrocks et al., 2004) est une extension du langage OWL qui permet d'exprimer des règles logiques sur les ontologies OWL en ajoutant une couche de logique déductive au modèle sémantique. SWRL permet d'écrire des règles pour faire des inférences que les axiomes ne sont pas capables

d'inférer. Ainsi, SWRL complète parfaitement les capacités de raisonnement logique d'une ontologie. Il sera présenté dans le prochain paragraphe.

1.3.3. Semantic Web Rule Language

Les règles SWRL sont des énoncés condition-action qui spécifient des conditions sous lesquelles des conclusions peuvent être tirées. Chaque règle SWRL est composée d'une partie antécédente qui décrit les conditions que les instances doivent satisfaire et d'une partie conséquente (action) qui spécifie les conclusion à tirer si les conditions de la partie antécédente sont satisfaites. Les conditions et les actions sont exprimées en utilisant des termes issus de l'ontologie OWL : (i) des instances de concepts C(?x) (i.e., ?x est l'instance du concept C) ; (ii) des relations entre deux instances R(?x,?y) (i.e., ?x et?y sont reliés par la relation R), (iii) des relations same-as(?x,?y) ou differentFrom(?x,?y) et (iv) des fonctions calculatoires (e.g., addition, supérieur à). Les seuls opérateurs logiques de ce langage sont des ET (^) (donc pas de OU et pas de SAUF).

Par exemple, a, b et c sont des instances de l'ontologie, si a Cause b et b Cause c alors a Cause de manière indirecte c.

Écrit de la façon suivante : Cause(?a,?b)^Cause(?b,?c) → Cause_indirecte(?a,?c).

Si une personne a plus de 18 ans c'est un adulte : Personne(?x), aPourÂge(?x, ?age), swrlb:greaterThan(?age, 18) → Adulte(?x)

Il est possible d'inclure une expression logique des deux côtés de la règle, c'est à dire qu'il peut exister plusieurs conséquences, par exemple : C(?a)^R(?a,?b)^R(?a,?c) → C(?b)^C(?c)).

1.4. Utiliser une ontologie

1.4.1.Les raisonneurs

Les raisonneurs, aussi appelés raisonneurs sémantiques ou moteurs d'inférence, permettent d'inférer les conséquences logiques d'un ou plusieurs axiomes ou règles d'inférence. Ils sont de plusieurs types et sont plus ou moins adaptés selon l'application de l'ontologie (e.g., nombre élevé de classes). Les trois plus utilisés sont Pellet (Sirin et al., 2007), Hermit (Shearer et al., 2008) et Fact++ (Tsarkov et al., 2006). Ces trois raisonneurs sont disponibles dans Protégé un éditeur d'ontologies (éditeur que j'utiliserai pour construire l'ontologie). Cet éditeur possède une fonction 'explain inference' qui permet à l'utilisateur de décortiquer le raisonnement qui a produit une inférence donnée. Les raisonneurs ne permettent pas de dialoguer à proprement parler avec l'ontologie, mais répondent

automatiquement à des questions formalisées sous forme de règles et d'axiomes. Les réponses aux questions posées sont représentées dans l'ontologie.

En biostatistique l'inférence consiste à extrapoler les résultats obtenus dans un échantillon à une population. En ontologie les inférences vont venir enrichir une représentation initiale par exemple en rajoutant des relations entre instances ou en catégorisant certaines instances en classes.

1.4.2. Les requêtes

Une requête correspond à une interrogation d'une base de données. De la même manière qu'avec des règles ou des axiomes, il est possible d'interroger l'ontologie que ce soit avant ou après les inférences produites par les raisonneurs.

Les modèles ontologiques peuvent être interrogés de différentes manières :

(i) Il est possible de considérer l'ontologie comme un graphe de connaissance. Par la suite, des opérations sur des graphes (e.g., quel est le plus court chemin entre x et y) peuvent être réalisées : par exemple, grâce un langage comme CYPHER intégré au système de gestion de base de données Neo4j (Technology, Inc. (2015). Neo4j, the World's Leading Graph Database. Neo4j Graph Database.).

(ii) Il est possible de réaliser des opérations d'algèbre relationnelle (e.g., selectionner les organes situés dans le haut du corps et vascularisés par x). Le langage de requêtes SPARQL pour SPARQL Protocol and RDF Query Language créé par le W3C permet d'interroger l'ontologie avec une requête de la forme SELECT, FROM, WHERE. Dans la majorité des cas, la clause FROM est omise car le graphe entier est parcouru et pas une sous partie. La clause SELECT permet d'annoncer quelle variable sera sélectionnée : un prédicat, un sujet, ou une valeur. La clause WHERE permet de préciser le type de variable que l'on souhaite sélectionner (e.g., les sujets avec un prédicat avec une valeur de X). Les requêtes SPARQL peuvent aussi renvoyer un nombre (COUNT) ou une moyenne (AVG). Plus récemment, un autre langage de requête appelé SQWRL (*Semantic Query-enhanced Web Rule Language*) basé sur les règles SWRL a été créé (O'Connor M et al., 2007). La requête permet de définir un antécédent et remplace l'expression logique du conséquent par l'équivalent du SELECT de SPARQL (e.g., C(?a)^R(?a,?b)^R(?a,?c) → sqwrl:select(?b,?c)). Avec le langage SQWRL il est aussi possible d'utiliser d'autres opérateurs tels que UNION, DIFF et INTERSECT pour réaliser des opérations ensemblistes.

(iii) Les règles SWRL et les axiomes sont intéressants pour inférer ce qui est représenté dans l'ontologie. Certaines informations ne sont pas forcément représentées dans une ontologie, par exemple, une variable causée par au moins trois variables et causant au moins deux variables. Pour ce faire, il est possible de réaliser des opérations logiques avec DL-query (*Description Logics query*) de la même manière que l'on rédige un axiome.

1.5. Les différents types d'ontologies

Les ontologies ont été typées de différentes manières. Par exemple, selon leur degré de formalisation (i.e., informelles à formelles), le but (e.g., l'application finale d'extraction de concepts de documents médicaux), la complexité (i.e., la granularité des classes), la généricité (cf. plus bas) ou encore le sujet ou domaine (e.g., la médecine). L'auteur Guarino (Guarino et al., 1998) définit quatre types d'ontologies que sont les *Top-level*, *Domain*, *Task* et *Application* ontologies. Les auteurs Gomez-Perez and Fernandez-Lopez (Gomez-Perez et al., 1999), (Gomez-Perez et al., 2003a) ajoutent 5 autres types (*knowledge representation*, *general* or *common ontology*, *linguistic ontology*, *domain-task ontology* et *method ontology*).

Afin de facilement situer la future ontologie sans se noyer dans des détails, seuls cinq types seront définis :

- *Top-level* or *upper-level* ontologies : les ontologies de haut niveau représentent des connaissances qui sont valables quel que soit le domaine de connaissance de l'ontologie. Ce sont des connaissances qui sont vraies pour tout le monde (indépendantes du domaine). Ce type d'ontologie est parfois dit *fundational*, car son but est d'être à la base (la fondation) de toutes nouvelles ontologies. Elle ne contient pas de classes d'un domaine donné, comme la médecine.
- Ontologies de domaine : elles posent un cadre conceptuel relatif à un domaine particulier, comme la médecine, ce qui permet de représenter des faits de façon formelle, à la manière d'un modèle de données.
- Ontologies de tâches : elles décrivent ce qu'est une tâche ou un processus de manière générale qui peuvent donc servir dans plusieurs domaines ou non. La tâche de sélection des variables pourrait être représentée sans pour autant que l'ontologie soit capable de sélectionner les variables.

Ontologies de domaine-tâches : ce sont des ontologies de tâches utilisables pour un domaine particulier, comme la description de la tâche de diagnostic en médecine. Elles sont construites de manière indépendante de leurs potentielles applications.

- Ontologies d'application : la construction de ces ontologies est guidée par l'application ou les applications. Elles doivent représenter toutes les connaissances nécessaires pour réaliser une tâche spécifique. La tâche en elle même n'étant pas représentée.

1.6. Méthodes de construction

Il existe plusieurs méthodes de construction des ontologies. Comme écrit dans Noy N et al., 2001 : *« There is no one correct way to model a domain, there are always viable alternatives. The best solution almost always depends on the application that you have in mind and the extensions that you anticipate. ».* Ces méthodes peuvent être semi-automatiques ou purement manuelles, elles peuvent partir du besoin initial ou d'un modèle ontologique déjà existant. Ci-dessous sont détaillées quelques unes des méthodologies :

Grüninger and Fox (Grüninger et al., 1995.) :

La méthodologie proposée consiste à définir plusieurs scénarios dans lesquels des problèmes se posent sous forme de questions auxquelles l'ontologie devra répondre (Scénario motivationnel et Question de compétence Informelle). Ensuite, la terminologie de l'ontologie sera spécifiée de façon formelle, ce qui permettra de définir les questions de manière formelle. Pour finir, les axiomes seront ajoutés et les questions de compétence testées.

Methontology (Fernández-López, M et al., 1997) :

Cette méthodologie considère la création d'une ontologie comme un processus itératif (« life cycle ») basé sur une succession de prototypes sur lesquels on peut modifier, ajouter, retirer des éléments, lors de chaque itération. Elle est découpée en **Spécification** (le but de l'ontologie, le niveau de formalité, le « scope » i.e., la liste des termes) ; **Acquisition** de la connaissance (source) ; **Conceptualisation** ; **Intégration** (utilisation d'autres ontologies) ; **Implémentation** ; **Évaluation** ; **Documentation**.

Ontology Development 101 guideline :

Ce guide de construction ontologique a été produit en 2001 par Noy and McGuiness. Il permet aux nouveaux utilisateurs et constructeurs d'ontologie de commencer rapidement à construire une ontologie en sept étapes : (i) Déterminer le domaine et le « scope » de l'ontologie : qu'est ce que nous allons faire avec l'ontologie, qui va utiliser et maintenir l'ontologie, à quel type de question de compétence l'ontologie doit elle pouvoir répondre ?; (ii) Considérer la réutilisation d'ontologies existantes ; (iii) Énumérer les termes importants de l'ontologie ; (iv) Définir les classes et la hiérarchie de classe : bas vers le haut, haut vers le bas, ou une combinaison (Uschold M et al., 1996); (v) définir les propriétés, (vi) définir les

contraintes des propriétés (e.g., types de valeurs autorisées, range et domaine) ; (vii) créer les instances. Il compile aussi quelques conseils ou erreurs à éviter (e.g., comment choisir entre instance et classe).

Il existe de nombreuses autres méthodes, telles que la méthode OTKM décrite dans (Sure Y et al., 2002) et (Sure Y et al., 2003), très inspirée de la méthode Methontology ou The Systematic Approach for Building Ontologies (SABiO) (Falbo, 2014) qui se focalisent sur une méthode de construction pour les ontologies de domaine. La future ontologie est une ontologie d'application et non de domaine. Présenter ces méthodes apporterait de la confusion quant au choix final.

Les méthodes basées sur l'extraction automatique de termes depuis des corpus de texte secondairement validés par l'humain ne seront pas développées dans cette thèse. (e.g., Archonte (ARCHitecture for ONTological Elaborating), Text2Onto, Terminae, OntoLearn Reloaded). Ces méthodes n'ont pas été utilisées pour la construction de la future ontologie et leur non utilisation est justifiée dans la discussion du Chapitre 4.

1.7. Éditer une ontologie

Il existe de nombreux éditeurs pour la construction d'ontologies. Ils peuvent être ouverts, gratuits ou non tels que Protégé (Musen MA, 2015), SWOOP (Kalyanpur A et al., 2006) et FluentEditor (https://www.cognitum.eu/semantics/FluentEditor/).

Aujourd'hui, l'outil principal de construction des ontologies et le plus accessible est le logiciel Protégé. Le développement de cet outil gratuit et ouvert a commencé en 1995 à l'université de Stanford. Il permet de manipuler tous les éléments constitutifs des ontologies (classes, propriétés, attributs, axiomes, règles SWRL), raisonner avec les raisonneurs décrits plus haut (Pellet, Hermit ou Fact++), et requêter l'ontologie avec SPARQL, SQWRL, ou DL-Query. La figure 12, présente l'interface graphique du logiciel Protégé.

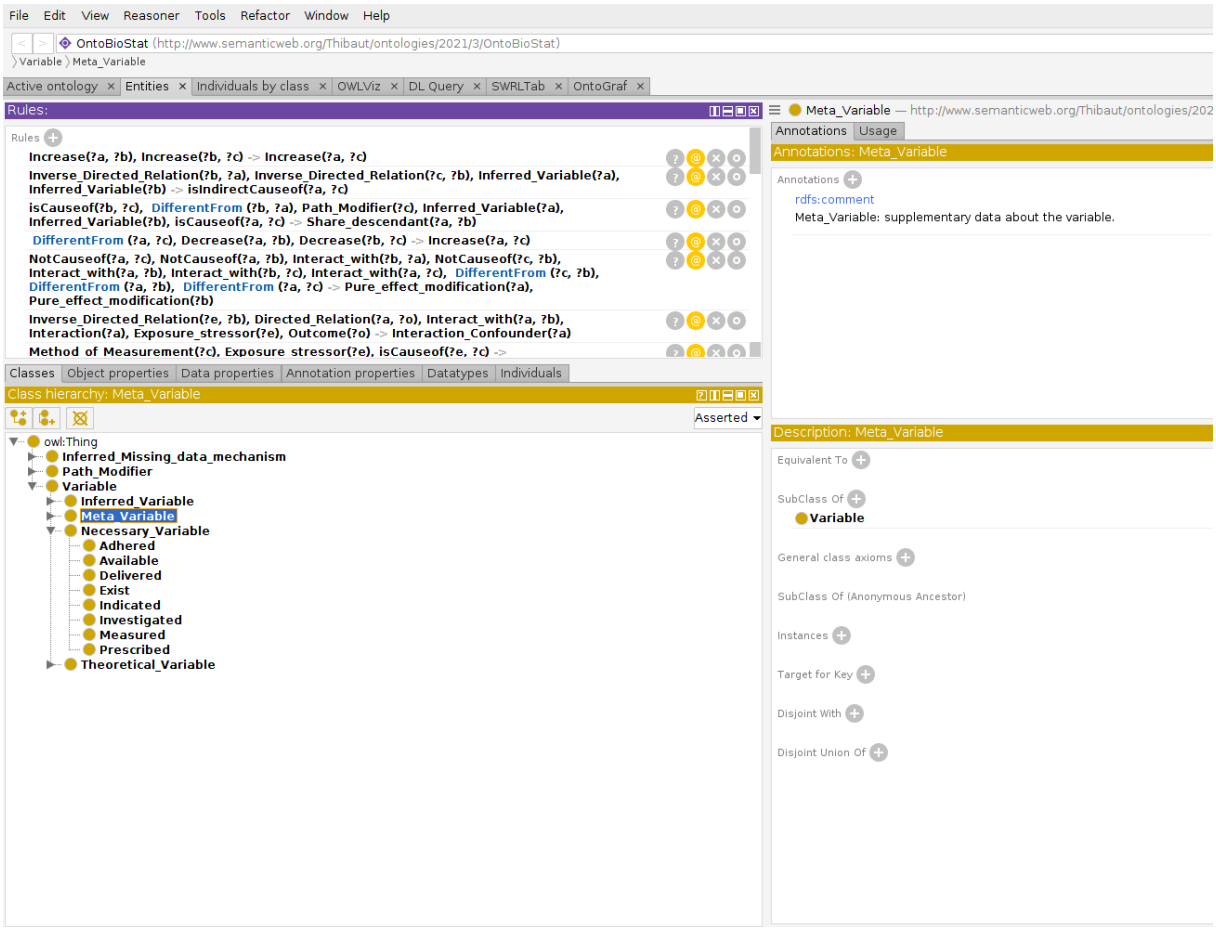


Figure 12. Interface graphique du logiciel Protégé.

1.8. Évaluation d'une ontologie

Afin de diffuser un outil ontologique pour l'utilisation d'un grand nombre de cas d'usage, il est indispensable de pouvoir s'assurer de sa validité. Il existe une liste importante de critères de validité (Gruber, T. R. et al., 1995 ; Gómez-Pérez, A. et al., 2001 ; Poveda-Villalón, M. et al., 2012) et de méthodes d'évaluation.

1.8.1. Critères de validation

Concernant les critères de validation, plusieurs auteurs ont proposé des listes qui se recoupent plus ou moins :

Clarté et objectivité (GRUBER). Il faut fournir une définition objective et une documentation en langage naturel. Ce critère est intriqué avec trois autres : (i) Concision (GOMEZ) : absence de redondances, que ce soit en terme de définition explicite ou inférentielle, qui sont des définitions non nécessaires ; (ii) Représentation du monde (POVEDA) : les définitions doivent correspondre à la réalité ; (iii) Compréhension humaine

(POVEDA) : l'ontologie doit fournir un maximum d'éléments pour être comprise par l'humain, tels que les labels ou commentaires, le format des noms ...etc.

Cohérence (GRUBER), **consistance** (GOMEZ) et **consistance logique** (POVEDA). Les inférences doivent être en accord avec les définitions. Un problème de consistance, ou inconsistance, correspond à des inférences contradictoires par rapport au cadre ontologique en lui-même, mais aussi en présence d'instances. Par exemple, la classe « Confounder » est forcément une sous-classe de « Covariable » ; une instance ne peut pas à la fois être Variable et Constante.

Extension possible (GRUBER), **évolutivité** et **sensibilité** (GOMEZ). Il faut que l'ontologie puisse évoluer (e.g., nouvelles classes) sans que cela n'entraîne une refonte trop importante de ses définitions ou de sa structure hiérarchique et sémantique.

Complétude (GOMEZ). Une ontologie est complète si elle représente avec exhaustivité ce qu'elle est censée représenter. C'est un critère difficile à évaluer et à comprendre en opposition avec l'hypothèse du monde ouvert des ontologies qui est incomplet (Tambassi, T et al., 2022).

Modélisation (POVEDA). L'ontologie est définie en utilisant tout le potentiel de modélisation mis à disposition (e.g., classe disjointe).

Spécification du langage ontologique (POVEDA). L'ontologie utilise bien le langage annoncé (OWL, RDF, ...etc).

Application sémantique (POVEDA). L'ontologie répond de manière adaptée aux problèmes qu'elle est censée résoudre.

1.8.2. Méthodes d'évaluation

Concernant les méthodes d'évaluation, il existe de nombreuses approches pour évaluer les ontologies, à la fois humaine (Gomez-Perez, 2004) et automatiques (i.e., raisonneur, OOPS) :

Concernant l'évaluation humaine : elle correspond à la validation sémantique des connaissances et celle du modèle dans sa globalité. Elle peut être réalisée par plusieurs personnes, à la fois des personnes du domaine modélisé et des ontologues. Des outils collaboratifs ont été développés, mais il est difficile de concilier une activité de correction de groupe qui mélange des individus avec des compétences hétérogènes en représentation de la connaissance.

Concernant l'évaluation automatique : elle regroupe des outils pour aider à la validation structurelle, syntaxique ou simplement de la consistance.

Les raisonneurs comme Pellet permettent de s'assurer de la consistance et de la satisfaisabilité des classes de l'ontologie. La satisfaisabilité des classes correspond au fait qu'une classe peut ou non avoir des instances. Si la consistance ou la satisfaisabilité ne sont pas respectées alors le raisonneur renverra un message d'inconsistance.

OOPS

ONTOLOGY PITFALL SCANNER ! (OOPS !) 5 (Poveda-Villalón et al., 2012) est un outil de validation qui s'utilise en ligne. Les auteurs ont défini une liste de pièges répertoriés en langage naturel dans un catalogue dont la majorité est détectée automatiquement: http://oops.linkeddata.es/catalogue.jsp. Ces pièges sont issus d'articles, de thèses et de livres sur l'évaluation et/ou la construction d'ontologies. L'ontologie peut être testée grâce au code source en RDF. Les résultats sont résumés sur une page avec une solution proposée en face de chaque piège retrouvé. Les pièges peuvent concerner des éléments individuels, plusieurs éléments ou toute l'ontologie. (Poveda-Villalón et al., 2014). Il est possible de regrouper les contrôles en cinq dimension : la structure comme par exemple P-17 « Overspecializing a hierarchy » qui consiste à définir une hiérarchie si profonde que les dernières classes ne peuvent pas avoir d'instance ; les relations par exemple P-03 création d'une relation *is* plutôt que d'utiliser subClass ou sameAs ; les noms et labels par exemple P-32 des classes différentes ont le même label ; les domains et les ranges par exemple P-11 : pas de domain ou range pour une relation ou un attribut ; et accessibilité de l'ontologie sur le web.

Pour finir, il est important de souligner que l'évaluation peut aussi reposer sur l'utilisation de l'ontologie dans un cas d'usage et la lecture de ses résultats (Porzel et al., 2004).

En conclusion de la partie évaluation, une ontologie doit : avoir une bonne syntaxe ; des termes non ambigus ; être cohérente du point de vue logique et est éprouvée par des tests ; être suffisamment expressive pour représenter les classes et relations ; être revue par des experts et des utilisateurs ; être suffisamment documentée ; fonctionner dans des cas d'usages réelles. L'interopérabilité n'est importante que dans certains cas où l'ontologie a vocation à être utilisée en conjonction avec d'autres.

1.9. Conclusion

Les ontologies sont une manière de représenter formellement la connaissance et de raisonner avec elle et les informations instanciées. Nous avons vu à travers divers exemples, comment elles pouvaient aider à représenter la connaissance de manière formelle avec un vocabulaire commun plus ou moins expressif, sous la forme d'un modèle de données, partageable, réutilisable et compréhensible par les machines. Dans la suite de cette partie, nous aborderons la revue de la littérature afin de trouver des ontologies sur le sujet de cette thèse.

Les classes et object property sont analogues aux nœuds et flèches des dags. *Dans un DAG, flèches et nœuds sont des conventions permettant de représenter les variables d'une étude. Cette représentation de variables provenant d'une étude équivaut à une instanciation. En effet, dans une ontologie, il faut distinguer la classe et son instance. Dans un DAG, il y a donc le nœud correspondant à la convention (classe) et le nœud qui est une variable réelle (l'instance de la convention). De même, une flèche, une fois instanciée, devient la relation entre deux entités spécifiques (par exemple, le lien entre le tabac et le cancer du poumon) au lieu de simplement deux nœuds.*

L'ontologie fournit une convention plus riche que celle des dags. *Elle propose donc des grandes familles (classes) de nœuds ou autrement dit des grandes familles de variables comme la famille des expositions ou celle qui regroupe les maladies. Dans chaque étude, une instance de la classe* ***exposition*** *sera retrouvée. Parfois, cette instance pourra aussi être l'instance de la classe* ***maladie****. Ces instances ou variables pourront être reliées entre elles par des flèches qui, dans le cadre de l'ontologie, seront des « object properties ». À l'instar des ontologies qui regroupent des nœuds en différentes familles de classes, on observe des regroupements de flèches en familles d'« object properties » (par exemple, cause de manière probabiliste versus cause de manière déterministe).*

En plus des nœuds et des flèches. *Les data properties, quant à elles, servent à relier une instance ou variable ou nœud à une valeur concrète (comme un pourcentage, une donnée littérale ou une catégorie). Il est ainsi possible d'indiquer le pourcentage de données manquantes d'une variable (« a pour proportion de données manquantes x % »).*

Pour ce qui concerne les raisonnements de l'ontologie à travers les axiomes et règles SWRL, *le point important est qu'on utilise une règle SWRL lorsqu'un axiome ne peut pas fournir un raisonnement suffisant. Le raisonnement se fait à partir de la représentation d'un graphe construit avec l'ontologie. Il se base donc sur l'appartenance des instances à des*

classes et/ou les liens (« object properties ») entre elles. Cela permet ensuite de déduire par exemple quelle variable est une variable de confusion.

2. Ontologies et sélection des variables dans l'inférence causale dans la recherche biomédicale : état de l'art

Le but de ma recherche est d'utiliser une ontologie d'application pour la sélection des variables afin de supprimer les biais pour estimer le vrai effet causal. Dans cette optique, il est nécessaire de faire un état de l'art des ontologies existantes sur le sujet à partir d'une recherche bibliographique.

2.1. Méthode

La recherche des articles s'est faite de manière itérative et agile. Les bases MEDLINE, Digital Bibliography & Library Project (DBLP) ont été utilisées. MEDLINE est le corpus d'articles scientifiques en biologie et médecine le plus utilisé dans le monde et servi par son moteur de recherche PubMed. DBLP publie un ensemble d'articles relatif au domaine de l'informatique. Lors de la construction de l'ontologie et son dépôt sur Bioportal, les alignements proposés avec certaines classes ont permis de retrouver d'autres ontologies (bioportal est un portail d'ontologies qui regroupe des ressources relatives au domaine biomédical). De plus, lors de la lecture des articles sélectionnés de nombreuses ontologies sont citées et ont pu être considérées.

La requête : pour MEDLINE, le choix des mots clés a été orienté par leur présence dans le thesaurus MeSH (Medical Subject Heading) puisqu'il est utilisé dans l'indexation des articles scientifiques contenus dans MEDLINE.

Ontologie : considérant que le mot clé Ontology est très important, il a été recherché dans le titre. En effet, dans un article traitant d'une ontologie (au moins une fois), le mot complet ou partiel (Onto) devrait apparaître dans le titre. Afin d'éviter une grande quantité de bruit, il a été fait le choix de recherche « onto* » dans le titre. Dans le domaine biomédical, l'ontologie appelée Gene Ontology est très connue et de nombreux articles traitent de celle-ci. Afin de diminuer encore le bruit, tout article avec Gene Ontology en terme MeSH a été exclu.

Sélection des variables : le mot « variables » dans « sélectionner des variables pour la construction d'un modèle » a plusieurs synonymes : Covariate, Covariable, Feature.

Causalité : une relation causale est parfois appelée relation étiologique en épidémiologie.

La requête suivante a été utilisée dans PubMed et son équivalent dans DBLP :

("Covariable selection" OR "Covariate selection" OR "Feature Selection" OR "Variable Selection") AND ("Causal" OR "Etiolog*") AND "Ontolo*"[title] NOT "Gene Ontology"[mesh]

Étant donné l'absence de résultats, une recherche par morceaux a été réalisée. Chaque requête comprenait systématiquement le mot clé sur les ontologies (Ontolo*) associées avec un terme et ses synonymes de la liste suivante : Causal OR Etiolog*, Variable OR Covariable OR Feature OR Covariate, Epidemio*, Confound*, Statist* OR Biostat*.

La pertinence des articles à conserver a été définie par la présence d'une ontologie en lien avec la représentation des connaissances pour la sélection des variables dans un cadre causal. En l'absence d'une telle ontologie, les ontologies de domaine généralistes comme des ontologies sur la recherche biomédicale, l'épidémiologie, la causalité ou les statistiques ont été considérées comme pertinentes alors que les ontologies de domaine médical précis ou de tâches comme une maladie donnée ou les procédures en chirurgie ont été écartées.

Les articles ont été regroupés en trois thématiques : Causalité, Statistique, Épidémiologie (cette thématique ne dépendait pas des mots clé qui avaient permis de trouver les articles mais du contenu). Certains articles pouvaient appartenir à plusieurs thématiques. Pour chaque ontologie, une description succincte avec l'objectif de l'ontologie ainsi que d'éventuels cas d'usage relatifs ont été extraits des articles scientifiques. Le but n'est pas de s'appesantir sur le type d'ontologie, ni le nombre de concepts. Dans la discussion seront abordés les potentiels points communs avec la future ontologie ou les bases sur lesquelles il sera possible de s'appuyer pour sa construction.

La dernière mise à jour a été effectuée le 30 septembre 2022 pour Pubmed et le 15 décembre 2022 pour DBLP.

2.2. Résultats

Termes +/- synonymes	Statistic	Epidemio-logy	Confoun-der	Variable	Causality
Nombre de références PubMed/DBLP	518/316*	144/8	7/0	149/316*	106/40
Nombre de références retenues	1/3	2/0	0/0	0/3	3/2

*Figure 13: Diagramme de flux de la recherche bibliographique sur Pubmed et DBLP. * Recherche concomitante des mots clé statistique et variable sur DBLP*

Un total de 11 références sont issues de la recherche bibliographique. BioPortal a permis d'identifier trois autres ontologies (SCO, RO, STATO). OBI a été présentée car faisant partie d'autres ontologies retrouvées telles que OBCS.

2.2.1.Épidémiologie

Epidemiology Ontology (EPO)

L'objectif est de pouvoir annoter des ressources produites sur le thème de l'épidémiologie à partir d'une plateforme qui partage des ressources relatives à l'épidémiologie. Elle permet d'aider dans les requêtes et à l'intégration de données épidémiologiques [Pesquita C et al., 2014].

Les classes peuvent correspondre à des indicateurs comme le taux de reproduction qui viennent compléter les classes des autres ontologies telle que IDO (infection disease ontology) (Cowell L et al., 2010).

Dans l'exemple donné, l'ontologie fait partie d'un annotateur manuel de documents qui permet de rechercher le bon terme avec la bonne définition au sein de l'ontologie (e.g., 'net reproduction rate' versus 'net reproductive rate').

Ontology of Biological and Clinical Statistics (OBCS)

L'objectifs de Ontology of Biological and Clinical Statistics (OBCS) (Zheng J et al., 2016) est de fournir une documentation statistique des études suffisante.

OBCS représente tout ce qui est en rapport avec les statistiques d'une étude, de la collection des données à la conclusion de l'analyse en passant par la visualisation des données. Cette ontologie contient des classes comme 'data collection' ou 'inferential statistical analysis' et regroupe d'autres ontologies comme OBI (Bandrowski A et al., 2016) et Information Artifact Ontology (IAO) (Ceusters W et al., 2012).

Dans un des cas d'usage, OBCS permet de représenter la partie statistique d'une étude que IAO et OBI ne représentaient pas :

IAO:**data set** *has specified input* OBI: **survival analysis data transformation** *is specified output of* OBCS : **confidence interval**.

Ontology of Clinical Research (OCRe)

L'objectif ambitieux de OCRe [Sim I et al., 2014] est de pouvoir aider dans la revue de la littérature et l'interprétation d'études précédentes pour poser la question de recherche, développer une nouvelle étude, mener l'étude, rapporter les résultats, interpréter et appliquer les résultats en pratique.

Les classes de OCRe peuvent par exemple correspondre à des éléments du schéma de l'étude (*design*) comme **Allocation scheme** ou **Allocation concealment method value**. Il n'existe pas encore de cas d'usage.

Public Health Ontology (*PHONT)*

PHONT fait partie d'une suite d'ontologies développées autour du Population Health Record qui retrouve et intègre différentes sources de données pour la publication d'indicateurs (*Okhmatovskaia A et al., 2014)*.

Objectif : cette ontologie a été développée pour encoder la causalité en épidémiologie.

Selon les auteurs cette causalité : (i) est probabiliste ; (ii) ne peut être additionnée ou comparée ; (iii) les effets bénéfiques et défavorables du même déterminant sur un problème de santé ne sont pas mutuellement exclusifs ; et finalement (iv) est individuelle, qui peut ou non être observée en tant qu'association statistique au niveau de la population.

Résultats : la relation causale entre cause et effet est scindée en deux sous la forme suivante : A causes some (Disposition and has_realization only B). Les auteurs donnent l'exemple de l'obésité et du diabète avec :

Obesity has_positive_effect_on some RiskOfDiabetes
RiskOfDiabetes ≡ Disposition and is_realized_in only OnsetOfDiabetes
OnsetOfDiabetes ≡ Process and results_in some DiabetesMellitus

Ici, l'obésité augmente positivement le risque de diabète et c'est le risque de diabète qui est une disposition réalisée chez les personnes qui débutent un diabète. Cela permet de dire que l'obésité n'a été un éventuel risque que si l'instance a débuté un diabète.

Il est aussi possible de diviser une variable en plusieurs variables comme le BMI afin de pouvoir représenter les effets en U (par exemple, BMI très bas et BMI très haut ont un effet positif sur la mort). De la même façon les interactions peuvent être représentée.

Study Cohort Ontology (SCO)

L'objectif est d'encoder toutes les informations au sujet des participants d'une étude souvent résumées dans le premier tableau descriptif de la population de l'étude. Les classes représentent donc les variables généralement retrouvées dans ce tableau. SCO a été utilisée pour évaluer l'applicabilité et la généralisation des résultats d'une étude réalisée sur une population pour l'appliquer à une autre (Franklin JDS et al., 2020).

Ontology for Biomedical Investigations (OBI)

L'objectif est l'interoperabilité entre différentes études du domaine biomédical. Parmi les classes de OBI on retrouve par exemple **study design**. Un des cas d'usage a consisté à utiliser OBI pour annoter les données d'une banque de données biomédicales. Cela a permis de passer d'une documentation en plein texte à une documentation standard et expressive contenant des éléments de logique. Par exemple, les synonymes sont identifiables directement et les hiérarchies de termes sont contruites automatiquement (Bandrowski A et al., 2016).

2.2.2. Causalité

States, Processes and Events, and the Ontology of Causal Relations

L'objectif est de clarifier les relations ontologiques causales ou "causal-like" entre états, processus et évènements instances de classes.

L'auteur présente toutes les relations qui ne sont pas des causes mais qui sont proches telles que : *terminate*, *allow*, *initiate*, *perpetuate*, *maintain* à travers divers schémas impliquant des états, des évènements et des processus (Galton A, 2012) (Figure 14).

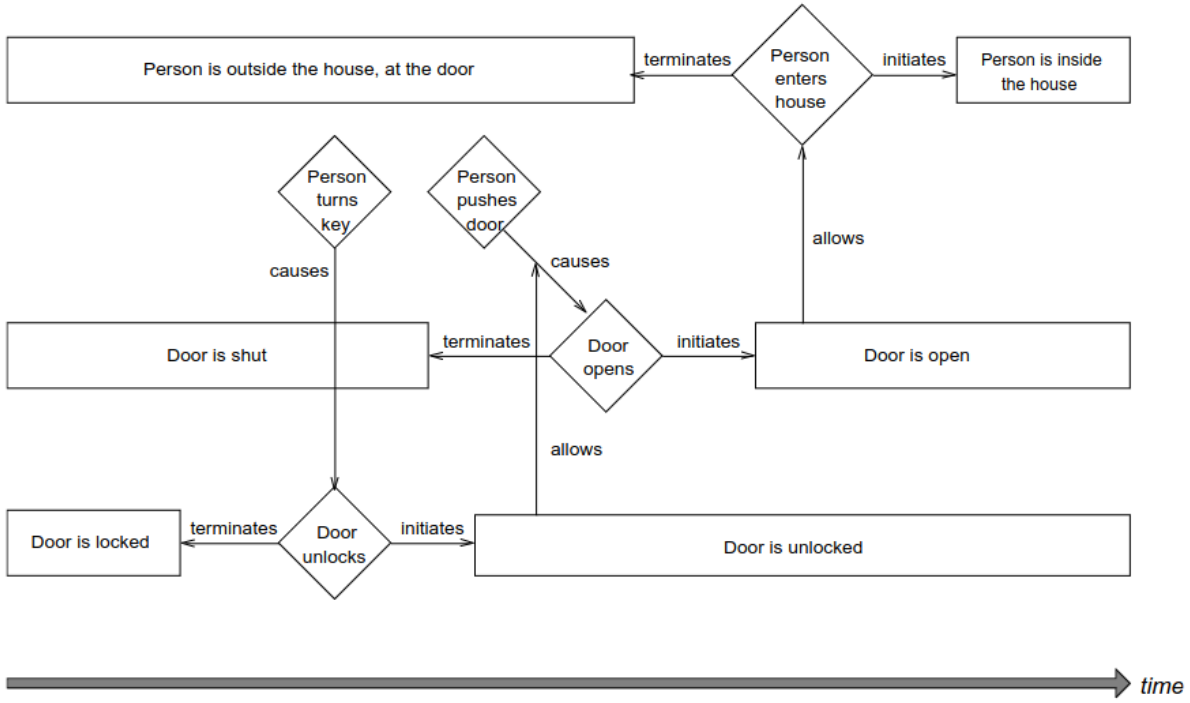


Figure 1. Causal and causal-like relations for a person entering a house.

Figure 14: Figure d'un exemple issue de l'article de Galton 2012

Dans cet exemple, pour pouvoir rentrer dans la maison la personne doit auparavant dévérouiller la porte, et la pousser afin qu'elle atteigne l'état de porte ouverte.

Radiology Gamut Ontology

L'objectif est de représenter les signes radiologiques et leur implication diagnostique. Radiology Gamut Ontology, est une ontologie dont les relations sont soit causales *(may_cause)* soit de subsomption *(is_a)*. Elles relient toutes les classes qui sont des signes d'imagerie et des diagnostics.

Il existe des règles de transitivité sur des relations de causalité *(may_cause)* (si A → B et B → C, alors A → C) qui tiennent compte de la subsomption *(is-a)* (Kahn CE, 2016*)*. Un de leur cas d'usage utilisait l'ontologie pour l'aide au diagnostic en radiologie (Kahn CE, 2014).

Gene Ontology Causal Activity Modeling (GOCAM)

GOCAM (Thomas PD et al., 2019) est une représentation qui se nourrit de la Gene Ontology qui est une ontologiereprésentant les gènes et leurs fonctions. GOCAM la remodélise en incluant des relations causales (*negatively regulates, enabled by*) ainsi que des éléments de contexte (*occurs in, happends during*) respectant le modèle suivant (figure 15). Cela permettrait de nouvelles applications de la Gene Ontology.

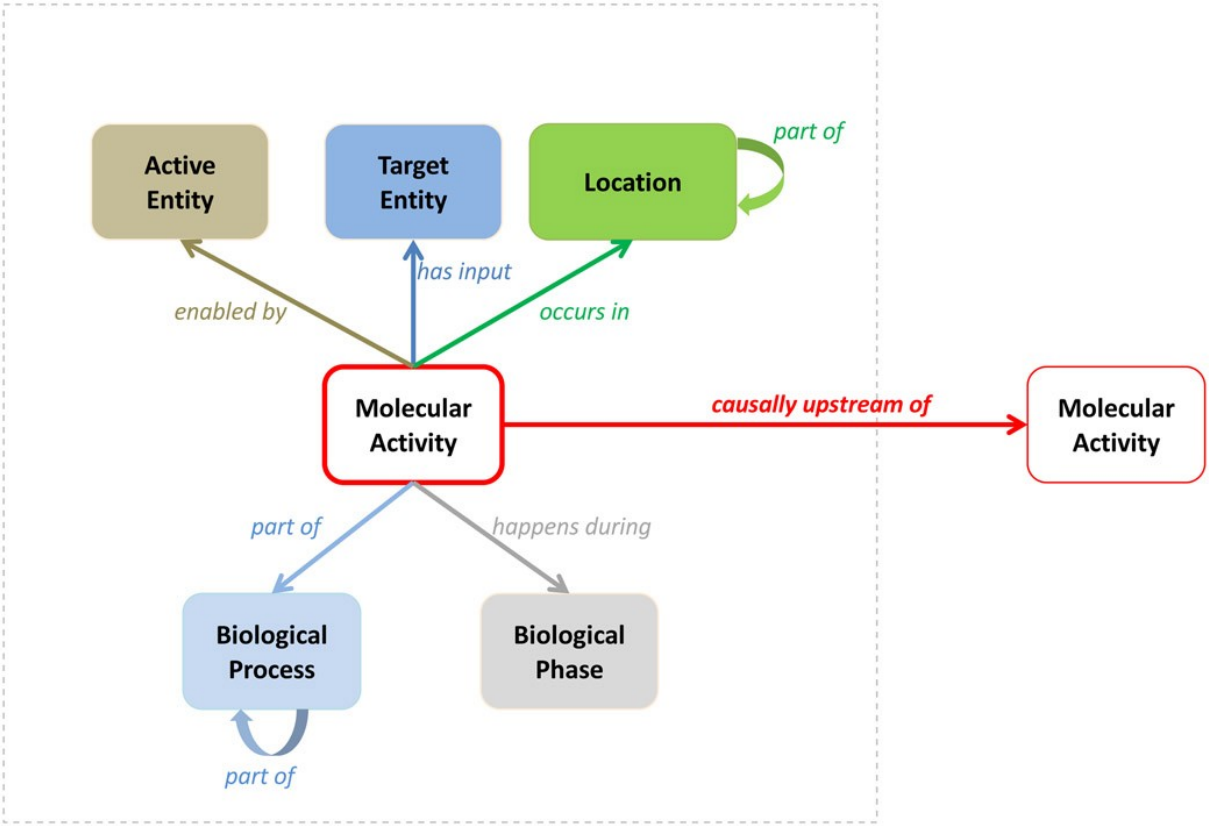


Figure 15. Gene Ontology Causal Activity Model extrait de l'article Thomas PD et al., 2019

Open Biomedical Ontology Relation Ontology (RO)

L'objectif est de formaliser toutes les relations du domaine médical pour aider les développeurs d'ontologies lors de la construction et l'annotation.

L'ontologie distingue trois types de relations : celles qui relient deux instances, deux classes ou une classe et une instance. La première version comportait des relations telles que *located_in*, *part_of*, *derives_from*, *has_agent*. Sa hiérarchie de relations compte maintenant plus de 641 relations (Smith B et al., 2005).

Ontology-Based Inference for Causal Explanation

Il ne s'agit pas d'une ontologie mais la définition d'un système de règles pour permettre des inférences. À partir des liens de causalité spécifiés, mais aussi de la hiérarchie is-a (e.g., un chat is-a félin), il est possible d'expliquer les relations entre deux classes ou instances. Par exemple, si A *is-a* B et B *cause* C alors A *cause* C. Ce système pourrait être utilisé dans les ontologies (Besnard Ph et al., 2008). Néanmoins, il n'a pas proposé les inférences utiles pour la sélection des variables à savoir l'inférence des ancêtres communs et conséquence commune.

2.2.3. Statistique et données

Statistics Ontology (STATO)

L'objectif était de représenter tous les temes utilisés en statistique que ce soit les tests statistiques ou encore les représentations graphiques.

À partir de cette représentation, il est possible d'annoter le contenu des articles scientifiques, d'améliorer le reporting dans les articles que ce soit au niveau des méthodes ou des résultats.

L'ontologie OBCS a intégré en partie quelques classes de STATO. Aucune publication scientifique présentant l'ontologie n'a été retrouvée (K. Kotis, A. Papasalouros, Statistics ontology, http://stato-ontology.org/, 2018)

StatsOnto

L'objectif est d'améliorer la transparence et la réutilisation des rapports d'analyse statistique.

StatsOnto encode jusqu'à la procédure statistique elle-même avec des classes telles que StatisticalTask et StatisticalMethod (Zheng et al., 2022).

Elle est capable de répondre aux questions de compétence suivantes : quelle est la structure des données d'entrée? et quelle méthode a t-on utilisé?

Statistical Learning Ontology (SLO)

Les objectifs sont : (i) d'établir une terminologie commune en lien avec le domaine de l'analyse de données; (ii) d'intégrer les connaissances d'un domaine et celles d'un jeu de données et (iii) d'utiliser ces connaissances exprimées dans l'ontologie pour assister différentes étapes telle que la sélection des variables.

Parmi les classes utilisées, il y a Model, Variable, Measure, LinkedVariables. Les relations entre variables ne sont pas représentées par des relations ontologiques mais des classes. Ces LinkedVariables peuvent avoir un LinkType (Causal, Hypothétique) et un LinkOrigin (viaModel, viaLiterature, fromExpert). Ces LinkedVariables ne permettent pas de préciser le sens de la relation (A → B ou B → A). Dans son modèle ontologique, il est possible de distinguer les variables indépendantes de la variable dépendante du modèles grâce au relation que la variable entretient avec la classe Model (*independantVariable* et *dependantVariable*).

Les questions de compétence génériques comprennent : Quelle est l'origine du lien entre variable 1 et variable 2 ? Comment mesure-t-on la variable X? Quelles sont les variables reliées à travers un modèle avec une autre variable? Quel type de lien existe entre deux variables?

Un des cas d'usages concerne correspond à la représentation de la connaissance du domaine du marketing digital. Dans ce cas d'usage la variable AdvertisingEffectiveness est reliée de manière causale aux variables Affiliation et SearchEngineScore (figure 16). Pour représenter ces relations SLO est obligé de créer plusieurs classes ou instances (non clairement expliqué) de la forme variable X *firstVariable* Link_1 *secondVariable* variable Y et Link_1 *linkType* Causal. Afin de répondre aux question de compétence, différentes requêtes SPARQL sont utilisées (Behnaz A et al., 2019).

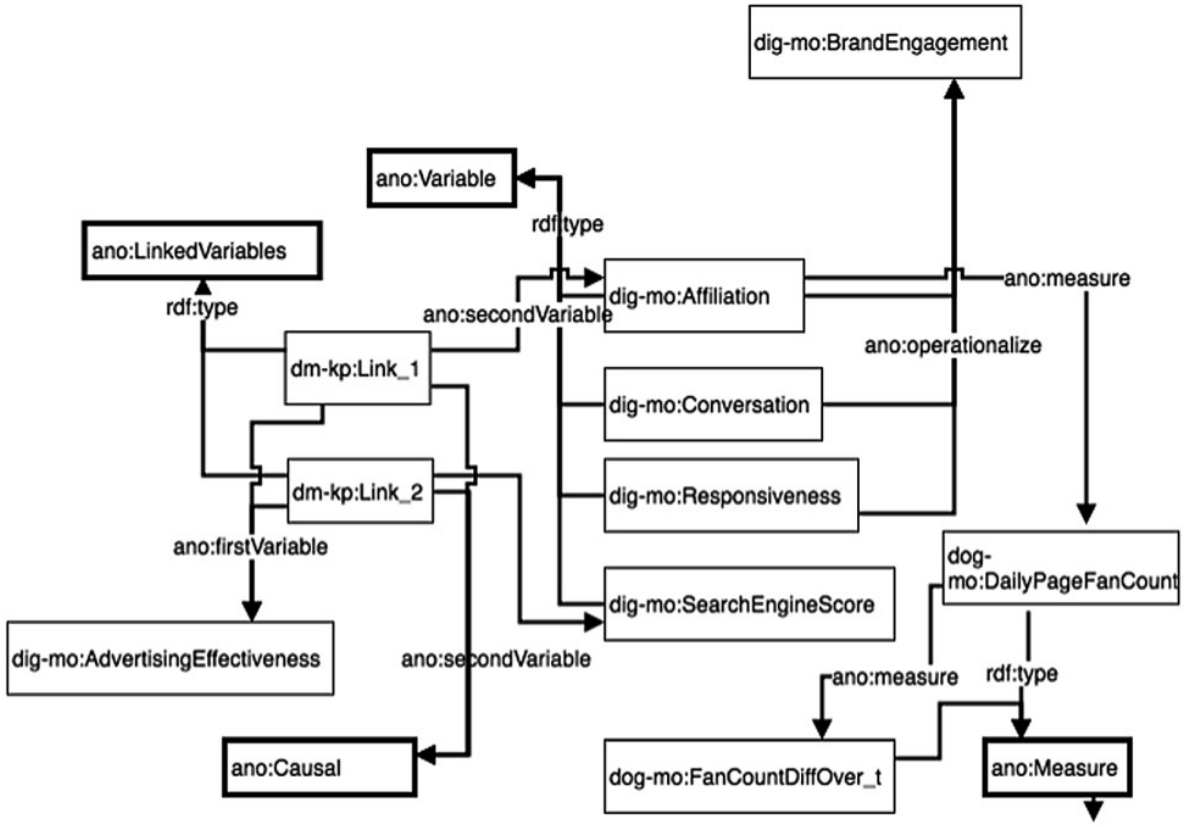


Figure 16: Cas d'usage de Statistical Learning Ontology : "Knowledge pack for digital marketing" (Behnaz A et al., 2019)

Dataset Characteristics and Quality Ontology (DCQ)

Le but de l'ontologie est d'aider les utilisateurs à choisir l'algorithme de sélection des variables pour leur modèle d'apprentissage machine (Nayak A et al., 2022).

DCQ représente les caractéristiques du jeu de données et la qualité de l'information que les auteurs regroupent sous le terme meta-feature. DCQ est composée de 39 classes telles que **Correlation** ou **Conciseness** (sous classe de Metrics). Il n'y a aucune représentation des relations entre variables.

Cette ontologie est capable de répondre à des questions de compétence telles que : en fonction d'un ensemble d'éléments à propos de la qualité de la donnée quel sera le meilleur algorithme de sélection ? Pour répondre à cette question elle s'appuie sur le nombre de valeurs aberrantes ou encore les corrélations existantes entre les variables associées à des règles SWRL. Les instances sont les jeux de données, les méta variables et finalement les algorithm de sélection des variables.

Pour l'instant, des tests ont été réalisées sur 10 jeux de données.

2.3. Conclusion

La sélection des variables est un processus complexe indispensable pour mesurer le vrai effet causal entre deux variables. Pour autant que l'on sache, aucune ontologie ne s'est intéressée à la représentation des connaissances dans le but de sélectionner les variables pour répondre à une question de recherche causale grâce à des inférences. Les deux ontologies qui s'en rapprochent le plus sont SLO et FSDCQ.

Actuellement, SLO ne fait aucune inférence. Elle permet donc uniquement de faire des requêtes sur les connaissances formulées par l'utilisateur via le langage SPARQL. Elle ne permet pas la sélection des variables dans un cadre causal pour la construction de modèle multivarié mais seulement la sélection de variables corrélées à d'autres pour diverses raisons (causale ou hypothèse, dépendante d'un résultat du modèle ou de l'opinion d'un expert). Le fait qu'elle représente les relations entre variables par une classe et non une relation dirigée rentre en conflit avec la représentation nœud, flèche dirigée des diagrammes causaux et les inférences potentielles qui découlent de cette représentation. Cette ontologie documente les relations entre variables qui pourraient avoir leur importance lors de la sélection des variables.

Concernant DCQ, la sélection des variables se fait en fonction d'informations dont on dispose sur le jeu de données. Par exemple, des informations statistiques qui nécessitent d'être calculées comme la corrélation moyenne entre chaque paire de variables d'un jeu de données. En revanche, il n'est pas fait mention de la connaissance *a priori* du lien entre les variables. DCQ n'est donc pas applicable à la sélection des variables basée sur les connaissances *a priori*. Cette ontologie, met en avant l'importance de la qualité des données, élément qui peut avoir son importance dans la sélection des variables.

Concernant la représentation de la causalité en médecine, PHONT est l'ontologie la plus poussée : Même si ce niveau de formalisme serait important dans les cas d'ontologies de domaine, les auteurs concèdent que pour des ontologies d'application cela a beaucoup moins

d’importance. En effet, affirmer que A cause B est très lourd avec PHONT à cause de la clause dispositionnelle. Bien que plus riche et détaillée pour représenter une relation causale, la clause dispositionnelle peut introduire une complexité qui rend délicate la tâche de définir une relation inverse ou de spécifier la nature transitive des liens causaux (la possibilité de chaînes de causes et d'effets). De plus, au niveau de la définition de la relation causale, les auteurs considèrent celle-ci comme INDIVIDUELLE alors que nous avons vu plus haut que les statistiques se basent sur des relations causales probabilistes, contrefactuelles et POPULATIONNELLES.

La suite de la thèse traitera du développement et de l’exploitation d’OntoBioStat, l’ontologie d’application que nous proposons pour la sélection des variables.

Chapitre 4 Développement d'OntoBioStat : une ontologie pour aider à la sélection des variables dans l'inférence causale

Dans ce chapitre, les différentes étapes de construction de l'ontologie OntoBioStat sont présentée. Dans le même temps chaque élément est décrit et contient des exemples. Enfin, dans la discussion, OntoBioStat est comparée aux DAGs et aux autres ontologies.

1. Matériel et Méthodes

La méthode de construction de OntoBioStat qui a été suivie correspond à une version modifiée de méthode de construction présentée par Natalya F. Noy and Deborah L. McGuinness dans Ontology Development 101: A Guide to Creating Your First Ontology : (i) domaine et cadre: but et applicabilité sont établis, (ii) réutilisation d'ontologies existantes, (iii) énumération des termes importants : vocabulaire compréhensible et partagé extrait de ressources qui font références, et (iv) création des classes, object properties, data properties : dans ce cas l'ontologie a été construite de manière incrémentale en commençant par les composants et les inférences de base des DAGs puis étendue en incorporant des composants plus avancés (e.g., interaction, flèche avec des signes positifs ou négatifs) et finalement l'addition des composants originaux de l'ontologie.

1.1. Domaine et cadre

Parmi les grands types d'ontologies décrits dans le chapitre 3 (1.5), OntoBioStat a été construite comme une ontologie d'application. Le domaine de OntoBioStat est la sélection des covariables (et le repérage des biais) pour l'inférence causale dans les études observationnelles en recherche biomédicale. La Figure 17 a été réalisée dans le but de situer la tâche que l'ontologie OntoBioStat est supposée supporter en utilisant des classes provenant d'autres ontologies sur la recherche biomédicale.

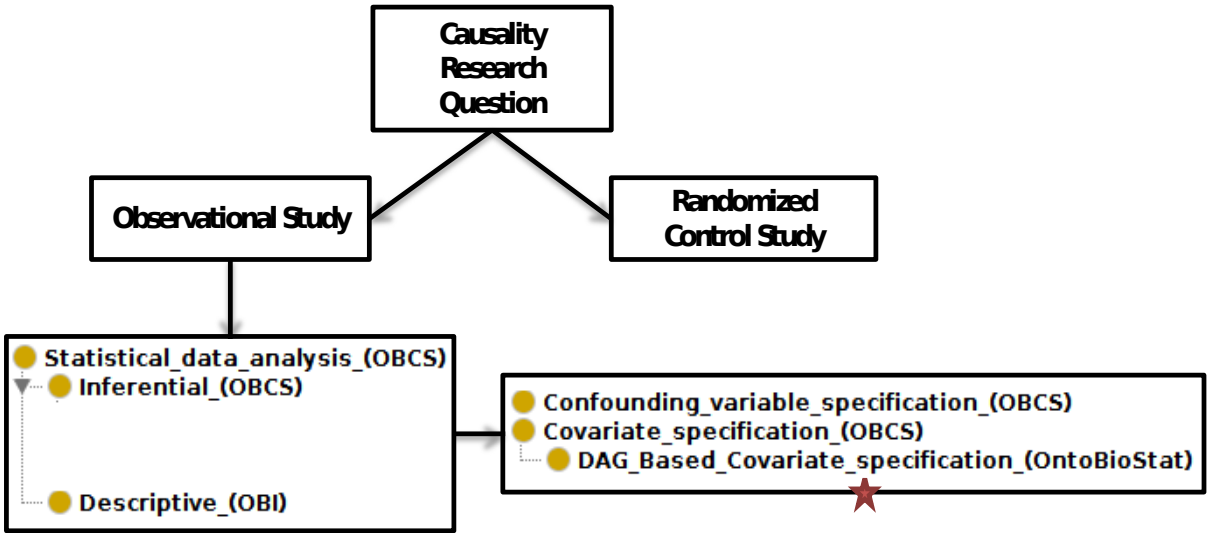


Figure 17: Quelle est l'application de OntoBioStat représentée avec des classes d'autres ontologies OBI: Ontology of Biomedical Investigation et OBCS: Ontology of Biological and Clinical Statistic

Il ne faut pas confondre une ontologie d'application avec une ontologie de domaine/tâche qui aurait représenté les acteurs ou encore les modèles statistiques ou les lieux de consultation de biostatistique. En tant qu'ontologie d'application, OntoBioStat ne représente pas non plus le processus de sélection des variables mais bien la façon de représenter les connaissances pour réaliser une bonne sélection de variables. Il est important de pouvoir définir : le périmètre de ce que l'ontologie peut représenter et les sources possibles de connaissances qui vont pouvoir nourrir l'ontologie. Le périmètre des connaissances nécessaires pour réaliser une bonne sélection des variables comprend : savoir théorique médical (e.g., guides de bonnes pratiques, ouvrages de médecine et de savoirs fondamentaux), connaissance pratique (e.g., protocole, équipement utilisé, équipe, période, lieu, etc.), et connaissance des données (e.g., le type de variable (quantitative ou qualitative), la base de données, les données manquantes, ...etc.). Il est indispensable que le périmètre englobant ces connaissances soit réaliste dans le sens où les classes et leurs potentielles instances, autrement dit les sources de connaissance, doivent être facilement accessibles pour le constructeur. Ces connaissances peuvent être récupérées dans le protocole d'une étude, la partie méthode, le rapport de la consultation du biostatisticien ou le jeu de données.

Après le domaine, il est nécessaire de spécifier quels sont les utilisateurs et quels sont les moments propices à l'utilisation de OntoBioStat. Les utilisateurs visés sont les utilisateurs des biostatistiques, incluant les novices et les cliniciens. L'ontologie pourra être utilisée lors de la rédaction de la méthodologie et aussi à la fin de l'étude (fin du recueil de données).

La sélection des variables comprend plusieurs étapes dans lesquelles l'ontologie peut aider : (i) la construction, (ii) l'analyse et la compréhension des diagrammes causaux.

Concernant la construction :

OntoBioStat doit se comporter comme un modèle de données pour un graphe de connaissances. Les flèches et les nœuds sont les composants primitifs des diagrammes causaux. Donc, la plupart des instances de classes sont des variables (nœuds) mesurées ou non. Ces instances sont connectées par des object properties correspondants à des flèches (e.g., relation causale).

Concernant les aides à la construction, rappels rapides de l'existant et de ses limites : Certains auteurs ont proposé des tutoriels généraux pour construire des DAGs. Les tutoriels ont été trouvés grâce à la requête PubMed sur les graphes orientés acycliques décrite plus bas. Les recommandations de construction mentionnent : des principes généraux tels que l'absence de flèche correspondant à l'absence de lien causal, les différences entre confusion, collision et médiation, ou encore le M-bias (Digitale JC et al., 2022) ; ou fournissent une méthode pour construire les DAG à partir de revue de la littérature sans pour autant proposer un cadre de construction formel et systématique pour le diagramme causal en lui-même (Ferguson KD et al., 2020). De plus, ces tutoriels manquent cruellement d'informations qui pourraient être utiles pour la sélection des variables telles que l'incertitude des liens causaux, le type de causalité (suffisante versus contrefactuelle) ou la qualité des données.

OntoBioStat a pour objectif de permettre la représentation d'éléments contextuels et d'autres éléments jusqu'alors implicites pour fournir un modèle riche et explicite dépassant ainsi la représentation classique d'un DAG et les conseils de construction basiques présents dans les tutoriels. De plus, cette construction se destine à être semi automatique grâce aux inférences.

Concernant l'analyse et la compréhension :

L'analyse d'un diagramme causal construit avec OntoBioStat doit permettre de répondre à une série de questions. Ces questions, sont appelées questions de compétences. Les questions de compétences sont complétées par des scénarios classiques en inférence causale qui adressent la réalisation d'analyse de sensibilité. Ces analyses de sensibilité ont pour but d'éprouver le résultat principal en faisant varier quelques paramètres et dans notre cas le jeu de variables à inclure. Ce qui pousserait à modifier le jeu de variables seraient la qualité des données, l'incertitude de liens causaux ou encore l'utilisation de variables dites substitues (proxies) de variables de confusion.

La création des questions de compétences non limitatives décrites dans le tableau 1 a été guidée par mon expérience personnelle en tant que biostatisticien méthodologiste. Ce sont

toutes des questions que se posent les biostatisticiens ou méthodologistes quand il faut sélectionner les variables dans un modèle ou quand ils se retrouvent face à un résultat statistique. Les questions de compétences même si inspirées de mon quotidien s'appuie sur des questions retrouvées dans les guides de lecture d'article. Elles représentent des questions fondamentales ou basiques qu'il faut se poser pour la sélection des variables.Elles peuvent être retrouvées dans des outils de lecture critique d'articles publiés par Cochrane, tels que ROBINS et ROB-2 (Sterne JA et al., 2016 ; Sterne JAC et al., 2019) ; Cochrane est un consortium pratiquant la méta-analyse. Leurs outils classent des articles en études à risque de biais majeurs versus études à risque de biais mineurs. Les outils ROBINS et le ROB-2 contiennent plusieurs dimensions dont certaines en rapport avec les biais de confusion, de mesure, de sélection, les problèmes de gestion des données manquantes et leur impact.

Tableau 1: Questions de compétences

1) Faut-il exclure des patients qui ne seraient pas comparable aux autres ?
2) Quelles sont les variables qui confondent le vrai effet causal entre exposition et critère de jugement ?
3) Existe-t-il une interaction qui pourrait biaiser le vrai effet causal entre exposition et critère de jugement ?
4) Est-ce que le mécanisme responsable de la présence de données manquantes pourrait biaiser l'estimation du vrai effet causal entre exposition et critère de jugement ?
5) Quelle est la direction des biais causés par des variable de confusion ?
6) Existe-t-il des variables proxies de variable de confusion ?
7) Quelle type de relation existe entre deux variables ?

Dans le tableau 1, différentes facettes de la sélection des variables sont présentes : (i) Exclusion de sujets, (ii) Sélection de variables de confusion, (iii) Interaction, (iv) Données manquantes, (v) Sens du biais, et (vi) Proxy.

Pour être sûr de pouvoir réaliser les bonnes inférences afin de mettre en évidence les potentielles variables de confusion, médiation et collision, il est indispensable de pouvoir définir d'un point de vue ontologique quelle relation unie deux variables (vii). Cette question a un double rôle puisqu'elle permet de définir si deux variables sont reliées, comment et pourquoi.

Il arrive que l'assignation du traitement ou de l'exposition de manière plus générale, frôle le déterminisme dans la population ou dans une sous population (contre indication ou indication absolue) (i). Or la causalité estimée dans un modèle stochastique est contrefactuelle, probabiliste et populationnelle et non déterministe, suffisante et individuelle.

Il n'est donc pas possible de corriger un tel biais. Il faut simplement exclure les sujets concernés. Néanmoins, il arrive qu'il n'y ait pas de division par 0 (référence au tableau de contingence entre la variable entraînant une assignation déterministe de l'exposition et l'exposition elle même). Ces sujets, en dépit d'une assignation a priori déterministe de l'exposition, n'ont pas été exposés. Ceux-ci doivent avoir des caractéristiques bien particulières qui, lors de notre tentative pour corriger le biais pourraient au contraire biaiser notre estimation. Cette problématique se reflète dans une pratique statistique qui consiste à retirer les patients avec une probabilité très élevée ou très basse d'être exposés, sachant leurs différentes caractéristiques (exclusion des sujets sur score de propension extrême).

Ensuite, il faut être capable de mettre en évidence les variables de confusion (iii) mais aussi les interaction (iv) qui pourraient confondre l'effet entre exposition et critère de jugement. En effet, deux variables sont corrélées si A cause B ou B cause A, A et B ont une cause commune, ou A et B ont une conséquence commune constante MAIS aussi dans le cadre d'une interaction. Si A cause l'exposition et que A interagit avec la relation entre la variable B et le critère de jugement, l'interaction AB confond l'effet causal entre l'exposition et le critère de jugement.

Pour juger de la qualité d'une variable il est primordial d'observer sa quantité de données manquantes et surtout le mécanisme sous-jacent responsable de l'absence de données. La question (iv) permet de déterminer si l'absence de données est complètement due au hasard ou non. Selon le mécanisme, le vrai effet causal peut être biaisé (Westreich D et al., 2012).

S'il n'est pas possible de sélectionner une variable de confusion de par son absence du jeu de données, il est alors possible de sélectionner une variable dite substitue (proxy) (vi).

Pour finir, en cas d'impossibilité totale de sélection d'une variable de confusion ou de sa ou ses substitues, il faut être capable de déterminer si cette variable biaise vers le haut ou vers le bas l'effet causal (v). Si sans ajustement, l'effet causal est sous-estimé cela permet d'affirmer que l'effet causal réel est au moins aussi important que l'effet causal estimé.

Une dernière question non présentée dans ce tableau a secondairement été rajoutée pour permettre de mettre en évidence des biais de suivi ou de mesure qui sont directement dépendants de l'exposition et la causalité inverse.

N'ayant pas fait l'objet d'un consensus d'experts, il est impossible d'affirmer leur complétude.

N’ayant pas fait l’objet d’un consensus d’experts, il est impossible d’affirmer leur complétude.

Le recodage des variables, qui fait partie de la sélection des variables, n’a pas été inclus dans cette liste de questions de compétence. Malgré l’intérêt certain de la représentation de la forme du lien causal entre deux variables, il n’y a pas de valeur ajoutée concernant les inférences. Cela signifie que l’utilisateur pourra le spécifier sans que cela n’entraîne d’inférence comme dans le cas d’une représentation en SKOS (Simple Knowledge Organization System) (Lénart M, 2007). Cependant, il n’est pas exclu que, dans le futur, l’ontologie puisse représenter le recodage.

Pour répondre à ces questions de compétences l’utilisation des axiomes et des règles SWRL a été prioritaire par rapport aux requêtes de types opérations sur graphes, ou l’utilisation de langages de requêtes SPARQL, SQWRL ou DL-query (chapitre 3 (1.4 : raisonneurs et requêtes)). En effet, nous considérons que les réponses aux questions de compétences sont des classes ou des object properties à part entière avec une définition propre. L’ontologie faisant partie des intelligences artificielles symboliques, il est possible d’obtenir les différentes étapes du raisonnement pour comprendre une inférence donnée.

En résumé, l’ontologie représente, d’une part, le cadre minimal pour la construction d’un graphe (i.e., l’instanciation des classes et des relations) et d’autre part, les classes et les relations provenant des inférences basées sur l’instanciation initiale qui répondront aux questions de compétences (analyse et compréhension).

1.2. Ontologies Existantes

L’utilisation de précédentes ontologies sur le sujet aurait pu permettre de démarrer sur une base existante. Pour autant que l’on sache, il n’existe aucune ontologie d’application dédiée au sujet de sélection des variables. Néanmoins, les termes d’autres ontologies pourraient être mis en correspondance avec notre ontologie.

1.3. Énumérer tous les termes importants

Un corpus de termes correspondant aux informations disponibles pour une étude donnée et compréhensibles par les utilisateurs des biostatistique a été créé. Ces termes ont été extraits d’articles de journaux publiés dans des revues à comité de lecture incluant :

- Les articles théoriques et pédagogiques sur le sujet des graphes orientés acycliques ou des diagrammes causaux pour la sélection des variables et le repérage des biais en

épidémiologie pour les études observationnelles. La requête suivante a été utilisée dans PubMed : *Directed Acyclic Graph* OR *Causal Diagram* with English filter. La requête a été décomposée de la manière suivante par PubMed :
(("direct"[All Fields] OR "directed"[All Fields] OR "directing"[All Fields] OR "direction"[All Fields] OR "directional"[All Fields] OR "directions"[All Fields] OR "directivities"[All Fields] OR "directivity"[All Fields] OR "directs"[All Fields]) AND "Acyclic"[All Fields] AND ("graph"[All Fields] OR "graph s"[All Fields] OR "graphed"[All Fields] OR "graphing"[All Fields] OR "graphs"[All Fields])) OR (("causal"[All Fields] OR "causality"[MeSH Terms] OR "causality"[All Fields] OR "causalities"[All Fields] OR "causally"[All Fields] OR "etiology"[MeSH Subheading] OR "etiology"[All Fields]) AND ("diagram"[All Fields] OR "diagrammed"[All Fields] OR "diagramming"[All Fields] OR "diagrams"[All Fields])).
La sélection des articles s'est faite à partir du titre, les extraits de PubMed, les résumés et si besoin le texte en entier. Les critères d'inclusion étaient : articles devant traiter du sujet de la sélection des variables ou les biais en épidémiologie pour les modèles statistiques et devant utiliser des graphes orientés acycliques ou diagrammes causaux pour représenter leurs propos. Les critères d'exclusion étaient : les articles des journaux cliniques et non biostatistiques ou non épidémiologiques;

- Des guides de reporting (les plus connus et généralistes) du site Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network (https://www.equator-network.org/reporting-guidelines/) ont été utilisés. EQUATOR est un site qui répertorie tous les guides de bonne pratique de reporting (manière de rapporter les études) en recherche biomédicale. Ces guides sont tous publiés dans des revues à comité de lecture. Il existe des guides spécifiques pour les études randomisées contrôlées CONSORT (Consolidated Standards of Reporting Trials) (Schulz KF et al., 2010), les études observationnelles STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) (*Elm E von et al., 2007*) mais aussi des sous-types d'études comme les études observationnelles de pharmacoépidémiologie réalisées avec les bases médico-administratives (Langan SM et al., 2018) qui sont des bases avec une grande quantité de variables dont le but premier n'est pas la recherche, mais le financement du système de santé. Les guides de reporting sont intéressants, car ils définissent l'information minimale qui doit être rapportée dans les articles afin de comprendre mais aussi de reproduire la recherche. Même s'ils peuvent être incomplets sur certains points, ils constituent une base solide pour commencer à énumérer les informations nécessaires pour la sélection des variables. Les guides suivants ont été utilisés pour créer le corpus : STROBE, CONSORT, SPIRIT (Standard Protocol Items: Recommendations for

Interventional Trials) qui est un guide pour rédiger les protocoles de recherche (Chan AW, 2013), RECORD-PE (Langan SM et al., 2018), TIDieR (Template for Intervention Description Replication) pour les études interventionnelles telles que les essais cliniques randomisés (Hoffmann TC et al., 2014).

Les termes ont été retenus pour différentes raisons : les termes permettent de comprendre les graphes orientés acycliques dans le sens du vocabulaire employé pour la lecture et la construction ; les termes sont des informations importantes pour la sélection des variables ou la potentielle survenue de biais (e.g., les détails concernant une étude donnée comme le lieu ou la période). Les termes comme variable de confusion (confounder), critère de jugement (outcome), exposition (exposure) étaient des termes déjà bien connus et n'ont donc pas été collectés à nouveau. Cependant, si des synonymes, antonymes ou des précisions (e.g., « time varying confounder ») sont trouvés, ils ont été ajoutés à la liste de termes. Les méthodes de sélection des variables basées sur les graphes orientés acycliques (e.g., back-door or disjunctive criterion) n'ont pas été retenues.

Chaque terme est défini par un extrait de l'article original ou une définition personnelle, ensuite au moins un lien vers une ressource a été fournie sous la forme d'un identifiant PMID (PubMed Identifier). Les éventuels synonymes et/ou antonymes ont été ajoutés. Si cela était pertinent, le nom de la représentation graphique était ajouté. Concernant les termes extraits des guides de reporting, seuls les termes non explicites ont été définis.

Le corpus a été validé par trois biostatisticiens (dont moi, deux biostatisticiennes titulaires d'un doctorat de santé publique exerçant en institut Inserm pour l'une et structure hospitalo-universitaire pour l'autre) habitués à l'utilisation des graphes orientés acycliques dans les études observationnelles. Pour être valide, un terme doit être compréhensible et sa définition doit être suffisamment claire. Chaque expert a déterminé si les termes devaient être exclus car hors sujet, considérés comme doublons ou synonymes/antonymes et si le terme décrivait une représentation graphique. Chaque définition fournie a été étudiée afin qu'elle ne soit pas ambiguë. La résolution finale au sujet d'un terme était atteinte par consensus.

Ces listes de termes vont pouvoir servir de base pour la création de l'ontologie.

1.4. Création de l'ontologie

OntoBioStat est représentée en Ontology Web Language (OWL) (Bock et al.) et a été construite avec Protégé 5X. Le raisonneur Pellet (Sirin E et al., 2007) a été utilisé pour les inférences et pour vérifier la consistance.

Contrairement au guide en 7 étapes de Natalya F. Noy 2001, ici, les classes ont été construites en même temps que les objects properties et les règles. Cela a entraîné plusieurs

itérations au fur et à mesure de la construction pour adapter la construction aux spécifications finales.

Pour la construction d'OntoBioStat, une méthode dite de « bottum-up » a été utilisée. Les éléments suivants ont guidé la construction : d'une part, la théorie des DAGs (un nœud = une variable, une flèche = une relation causale), avec des représentations graphiques théoriques nombreuses (exemple en figure 18), et d'autre part, le corpus de termes ainsi que les questions de compétences. Concernant les trois premiers diagrammes représentés sur la figure 15, variables de médiation, collision et confusion ont déjà été définies. Le quatrième représente un diagramme causal enrichi par les classes de OntoBioStat qui sont **Direct_confounder** et **Confounder_like**. Les deux sont des variables de confusion mais la ligne de raisonnement qui mène à cette conclusion est différente.

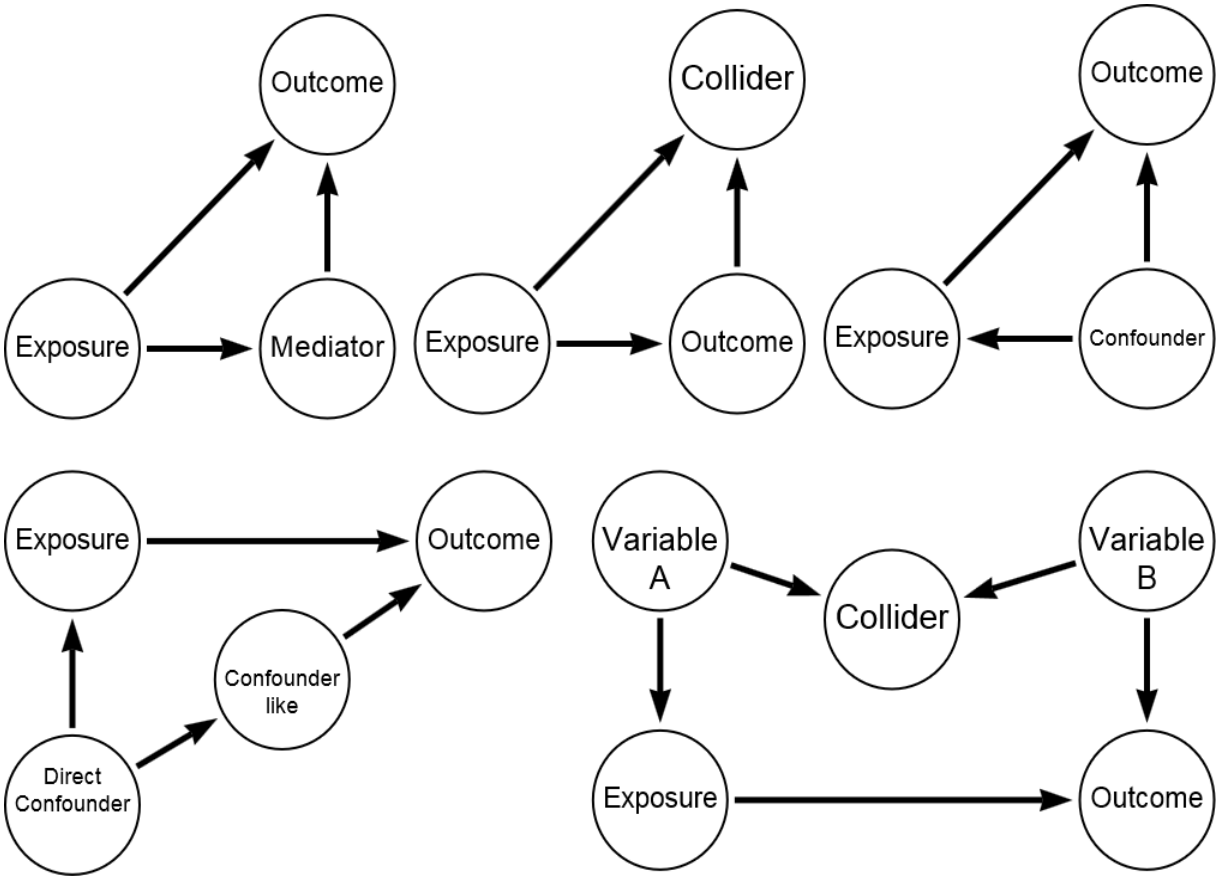


Figure 18. Exemples de DAGs et de classes d'OntoBioStat

Lors de création d'une ontologie, il n'est pas toujours évident de savoir quand préférer la création d'une classe à la création d'une data properties. Dans OntoBioStat, la création de data properties a été préférée à la place des classes lorsque pour une instance donnée il n'existe pas de variation (e.g., la variable A instance de Variable, a un data property

hasType égal à quantitative). Le domaine des data properties concerne uniquement les instances de variables.

Dans les ontologies, il existe des conventions de nommage. Dans OntoBioStat, le nom d'une classe commence avec une lettre capitale. Les classes inférées par les axiomes ou les règles SWRL ont le prefixe 'Inferred_', ou sont des sous-classes de ces classes (sauf exceptions).

La documentation est intégrée à l'ontologie. La définition en langage naturel (avec les éventuels synonymes), sources et justification(s) avec un ou plusieurs exemples ont été précisés en tant que commentaires. En tant qu'ontologie d'application pour la sélection des variables, la nécessité pour OntoBioStat de s'aligner avec d'autres ontologies ou de commencer par une ontologie de haut niveau (top level) n'est pas obligatoire. Contrairement aux autres ontologies présentées en amont, OntoBioStat ne sera pas utilisée pour la recherche d'information ou comme technique de traitement automatique de la langue, donc les synonymes ne sont pas indispensables à cette étape, même s'ils peuvent avoir leur importance pour la compréhension des utilisateurs. Dans ce travail, les classes d'OntoBioStat n'ont pas été mises en correspondance avec les classes d'autres ontologies. Cependant, si des classes similaires ou identiques existent, cette information a été spécifiée.

La dernière version de OntoBioStat est disponible en ligne sur BioPortal : https://bioportal.bioontology.org/ontologies/OBS. Le versionnage n'est pas pertinent dans notre cas, car seule la dernière version est utilisée.

Afin de fournir suffisamment d'informations concernant l'ontologie, la rédaction de la description de l'ontologie s'est appuyée sur le guide MIRO (Minimum Information for Reporting an Ontology). C'est un guide qui a été créé pour améliorer la complétude du reporting des ontologies ainsi que leur homogénéité (Matentzoglu N et al., 2018). Il énumère tous les éléments obligatoires et optionnels qui doivent être retrouvés lors de la présentation d'une ontologie.

1.5. Validation

Le corpus a été validé (validation sémantique) par trois professionnels habitués à l'utilisation des diagrammes causaux. Les articles scientifiques dont sont issus les termes permettent d'avoir une définition concise, objective, étendue au besoin du terme et compréhensible par l'humain. Quand de nouveaux termes ont été créés en dehors du corpus, leur définition en langage naturel ainsi que leur axiome, ou règles SWRL, correspondant permet de les comprendre.

Afin de s'assurer du maintien la cohérence/consistance après chaque modification des règles SWRL ou des axiomes, le raisonneur Pellet est testé sur plusieurs jeux d'instances théoriques et les éventuels cas d'usage. La validation par l'outil OOPS fait suite à la description de la construction de l'ontologie

Le chapitre 5 est dédié aux cas d'usage.

2. Résultats

2.1. Corpus

2.1.1. Termes issus des articles sur les DAGs

En septembre 2021 un total de 3.687 articles ont été retrouvés par la requête suivante: Directed Acyclic Graph OR Causal Diagram (Figure 19). Le nombre élevé de faux positifs facilement écartés grâce au titre ou nom du journal, vient de la requête étendue de PubMed incluant par exemple *diagrams*, *direct*, *graphs,* etc, dans tous les champs. Seulement 84 ont été initialement retenus basés sur le titre et les extraits. Trente-six articles sont secondairement exclus après la lecture du résumé et du texte en entier, 48 lus et 44 termes extraits. Après un consensus d'expert, quatre termes ont été exclus, quatre ont été considérés comme une représentation des diagrammes causaux, un comme un synonyme, trois comme antonymes, et un comme un doublon. Des définitions ont été rajoutées ou reformulées pour 13 termes (Tableau 2). Dans ce tableau, les termes sont regroupés par types : les DAGs signés, les interactions, la médiation, les biais et les données manquantes.

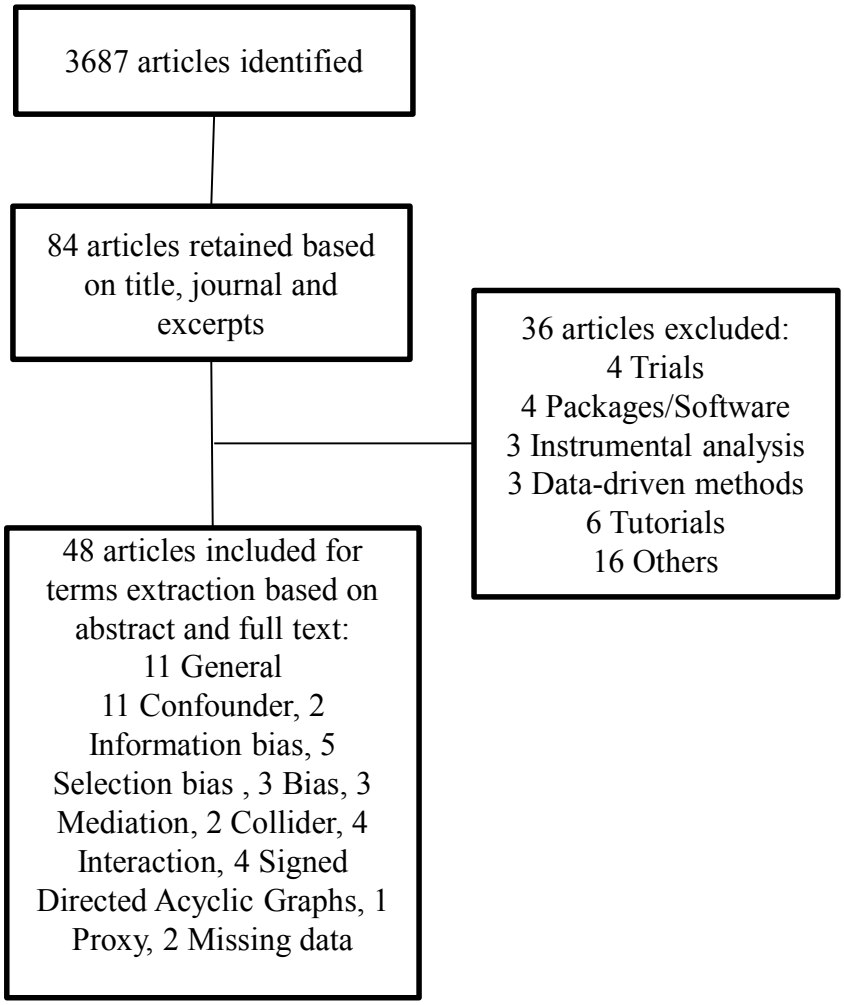


Figure 19. Diagramme de flux de la sélection d'article de MEDLINE : Directed Acyclic Graph ou Causal Diagram.

Tableau 2: Termes issus des articles sur les diagrammes causaux et les graphes orientés acycliques

Termes	Definition_synonymes	Sources	Synonyms	Antonyms	Representation
Common cause	a covariable which is ancestor of two or more variables	17702973.17700242.22904203.9888278	Share ancestor/parent, shared cause	bidirectional arrow	
Common effect	a covariable which is descendant of two or more variables	17702973.22081062.9888278	Share descendant, child	undirect or non directed edges/arc	
Descending proxy	variable A caused by variable B is a descending proxy of B	19525685		ascending proxy	
Uni-path collider	E0 and D1 collide at S, but the causal path from E0 to S passes only through D1.	23516121			
determinant of selection	affect exposure or outcome but never both	24598413			
time-varying confounding	Confounder that vary across time	20838848			
Monotonic effect relation	an exposure affect all persons the same way, intervention is always harmful or neutral for all individuals	17702973.18633331.25419168			
positive average montonic effect	the parent will increase or leave unchanged the average value of the child over the population	18633331		negative average montonic effect	signed edges
distributional monotonic effect	“In the case of a binary outcome Y all that is required for a weak monotonic effect is that a higher value of A makes the outcome Y at least as likely regardless of the value of the parents of Y other than A.”	18633331.25419168	weak positive monotonic effect		
Sufficient cause	MeSH definition: cause are	17702973			

Termes	Definition_synonymes	Sources	Synonyms	Antonyms	Representation
	sufficient when they initiate or produce an effect				
effect modification	effect of one variable on another varies across strata of a third	17700242.17700243.31121303	interaction,S-Variable, selection factor, effect measure modification	interact antagonistically	
X direct effect modifier	X is a direct cause of D and modifiesthe causal effect of E on D	17700242.17700243.33221880	direct interaction		
C indirect effect modifier	C is an indirect cause of D through X and modifies the causal effect of E on D	17700242.17700243.33221880	indirect interaction		
M effect modifier by common cause	C is a common cause of M and X, E and X direct cause of D, and M modifies the effect of E on D	17700242.17700243	share indirect cause		
R effect modifier by proxy	R is a proxy of X, direct cause of D, such as conditioning on R gives information on X and modifies the effect of E on D	17700242.17700243.33221880	confounded interaction		
pure effect modification	Given two variables A and B, not cause of a third variable C. Only the combination of A and B cause C.	17700243	causal co-action, 2-component sufficient cause		
total interaction	Direct + indirect effect modifier	33221880			
interact synergistically	effect of the combination is superior to the sum of the two effects	17702973			
Intermediate variable	an intermediate is a mediator	19525685.17702973.20838848.17700242	mediation variable, mediating variable		
indirect effect	that is, an effect mediated by another variable in the diagram	19330454.21556286			

Termes	Definition_synonymes	Sources	Synonyms	Antonyms	Representation
direct effect	that is, an effect not mediated by another variable in the diagram	19330454.21556286			
controlled direct effect	Given an outcome Y , treatment value of interest a, value settings m for some other observable variables, and value settings u to all background variables, the controlled direct effect of A on Y not via M is given by Ya,m(u) − Ya ∗ ,m(u).	19330454.21556286			
pure direct effect		19330454.21556286			
total effect	Sum of total direct effect and pure indirect effect	20838848.19330454 21556286			
natural direct effects	Corresponds to pure direct effect & total direct effect	19330454.21556286			
nondifferential	Concerning measurement error : Exposure true value or outcome true value is respectively independent of exposure or outcome measured	19755635.19366394		differential	
Dependent	The reason of measurement bias is the same for exposure than outcome. They share an ancestor	19755635		independant	
unmeasured	variable not collected or unavailable in the dataset	22904203	unobserved		
missing at random		21389091.30124749			
missing not at random		21389091.30124749			
missing compeltely at random		21389091.30124749			

2.1.2. Termes issus des guides de reporting

Un total de 41 termes (ou groupe de termes) ont été extraits et validés. Parmi eux, il y a des termes redondants comme Period qui se rapportent à la mesure de durée de différentes entités. Il y a aussi des termes qui se rapportent à une seule entité comme traitement (ever treatment, prior treatment, current treatment) mais qui pourront être appliqués à toutes sortes d'exposition (Tableau 3).

Tableau 3: Termes extraits des guides de reporting

Termes	Sources	Synonymes
Place	STROBE	Location
Period of recruitment	STROBE	
of exposure	STROBE	
of follow up	STROBE	
eligibility criteria	STROBE	
methods of selection	STROBE	
methods of follow up	STROBE	
effect modifier	STROBE	
diagnostic criteria	STROBE	
method of measurement	STROBE	
quantitative variable	STROBE	
number of participants with NA	STROBE	
category boundaries for continuous variable	STROBE	
data sources	STROBE	
direction of potential bias	STROBE	
sources of potential bias or imprecision	STROBE	
time window	RECORDPE	
current treatment	RECORDPE	
prior treatment	RECORDPE	
ever treatment	RECORDPE	
cumulative treatment	RECORDPE	
lag period	RECORDPE	
induction period	RECORDPE	
unexposed period	RECORDPE	Grace period
confounding by indication_contraindication	RECORDPE	
prescription start and end	RECORDPE	

Termes	Sources	Synonymes
new user	RECORDPE	
prevalent user	RECORDPE	
naive new user	RECORDPE	
how (intervention, outcome)	SPIRIT	
when (intervention, outcome)	SPIRIT	
criteria discontinuing/modifying	SPIRIT	
adherence	SPIRIT	
concomitant care_intervention	SPIRIT	
intervention provider	TIDIER	
modes delivery	TIDIER	
duration treatment	TIDIER	
intensity treatment	TIDIER	
dose treatment	TIDIER	
personalised intervention	TIDIER	
necessary infrastructure	TIDIER	

2.2. Développement de l'ontologie

D'abord, la démarche de construction de l'ontologie sera présentée pas à pas, ensuite un résumé de l'ontologie sera fourni. La démarche de construction de l'ontologie présentée ci-dessous s'attachera à représenter ce que les DAGs ou CDs peuvent déjà représenter ou inférer. Des représentations originales seront ajoutées soit au fur et à mesure, soit après la présentation des entités relatives aux DAGs ou CDs.

Les diagrammes causaux sont constitués de deux éléments principaux : les variables et les relations causales. Celles-ci permettent de déduire ou d'inférer si les variables sont à sélectionner ou non.

2.2.1. Un cadre de construction minimal

Les variables

Concernant les variables, il est nécessaire de distinguer celles qui sont réellement variables et les variables qui modifient le chemin causal (**Path_Modifier**). En effet, comme mentionné dans le chapitre sur les diagrammes causaux, un biais peut naître d'une variable qui est constante dans un échantillon donné. Pour représenter le biais de Berkson ou *Collider bias*, il faut donc représenter ces variables qui peuvent être constantes selon les études. Nous avons vu qu'un biais peut être corrigé ou créé grâce aux méthodes d'appariement,

d'ajustement, de stratification, d'exclusion (variable constante) ou de pondération. Ces méthodes transforment les variables en **Path_Modifier**. Les deux premières classes disjointes de l'ontologie OntoBioStat sont donc **Path_Modifier** et **Variable**. Les sous classes de **Path_Modifier** sont les méthodes citées plus haut : **Matched, Stratified, Adjusted, Ponderated, Excluded**. Les critères d'éligibilité demandés dans les guides de reporting peuvent être représentés via un **Path_Modifier** (e.g., uniquement les patients hospitalisés).

Dans l'inférence causale on s'interroge sur l'existence d'un lien de cause à effet entre exposition et critère de jugement. L'ontologie, doit donc distinguer le critère de jugement, l'exposition et les variables candidates pour la sélection, autrement dit, les covariables. Les trois classes disjointes **Outcome**, **Exposure** et **Covariate** sont ajoutées en tant que SubClass de **Variable** (figure 20).

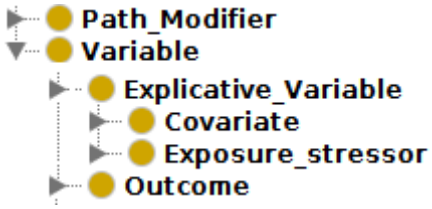


Figure 20: Hiérarchie des classes initiales

Afin de résoudre les algorithmes de sélection des variables à partir des diagrammes causaux, il est nécessaire de savoir si une variable donnée a été collectée ou mesurée. Dans le cas où le statut mesurée n'était pas spécifié, une covariable pourrait être considérée candidate pour la sélection alors qu'il n'est pas possible de la sélectionner. C'est à dire que l'algorithme proposerait de la sélectionner alors qu'il n'y a aucune données concernant cette variable. Cette propriété ne peut prendre que deux valeurs : mesurées ou non mesurées. La datatype property 'Measured' (TRUE/FALSE) est créée.

Les relations causales

La lecture des différents articles a permis d'isoler le vocabulaire de base des relations utilisées dans les diagrammes causaux. Ces relations ont été représentées sous la forme d'object properties (Tableau 4) et de nouvelles relations ont été créées : les relations inverses et les relations indirectes (figure 21). L'object property Share_descendant correspond au cas où deux variables ont une conséquence commune qui est un Path_Modifier.

Tableau 4: Récapitulatif des relations existantes et des nouvelles relations

CD Nom	Nom de la Représentation dans les CDs	Représentation	Ontologie	Properties
Direct effect/child/descendant	unidirectional/ single headed arrow/arc/edge	A → B	*isCauseof*	Asymetric, irreflexive
			hasCause	Inverseof isCauseof Asymetric, irreflexive
Indirect effect		A → B → C	*IndirectCauseof*	Asymetric, irreflexive
			hasIndirectCause	Inverseof IndirectCauseof Asymetric, irreflexive
common/ sharecause/ ancestor/parent	bidirectional/ two-headed arrow	A ↔ B	*Share_ancestor*	Symmetric, irreflexive
		A ←C↔D→ B	*IndirectShare_ancestor*	Symmetric, irreflexive
Share descendant	undirected/non-directional path	A →D - - B	*Share_descendant*	Symmetric, irreflexive
			IndirectShare_descendant	Symmetric, irreflexive

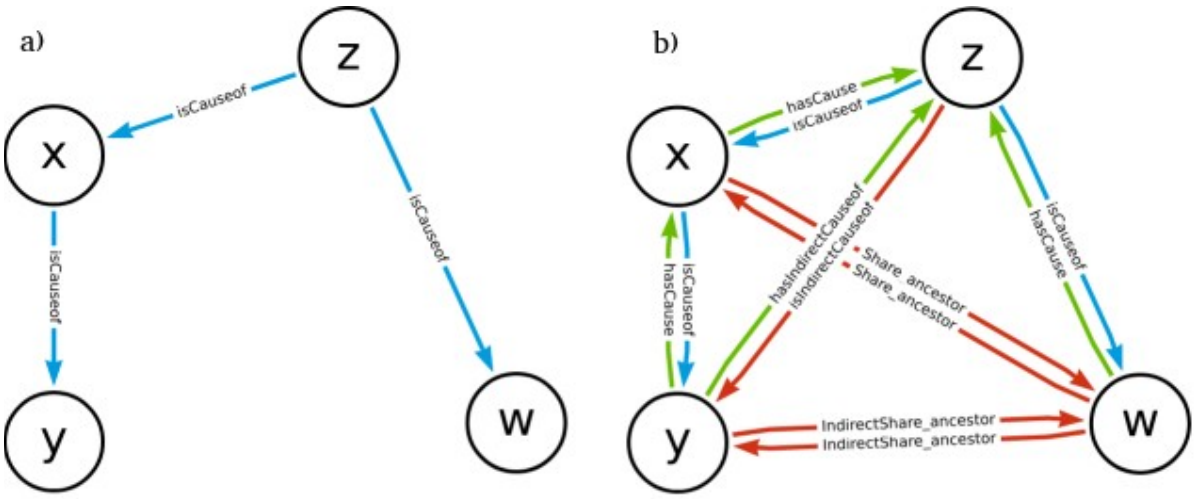


Figure 21. relation avant (a) et après (b) inférence avec les nouvelles relations : relations inférées à partir de règles (en rouge) et relations inférées grâce à la propriété inverseof en vert.

L'étape suivante est d'inférer les relations à partir de règles SWRL :

Les informations utilisées par les règles seront limitées à ce qui suit :

- Les règles devront être capables d'inférer à partir de relation *isCauseof* uniquement.

- Dans toutes les règles créées x, y et z sont des individus différents dans le but de respecter la propriété non réflexive (irréflexive) des relations. En effet, soient 4 variables w, x, y, z: z *isCauseof* x et w, et le couple (x, w) *isCauseof* y (figure 22). Sans la spécification 'differentFrom' y *Indirect_Share_ancestor* avec lui même. Les règles présentées ne contiennent pas la mention 'differentFrom' afin de conserver une certaine lisibilité de celles-ci.

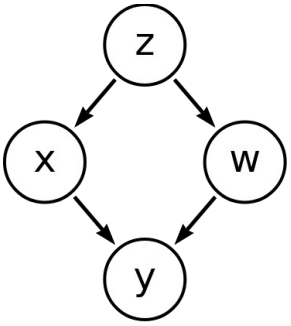


Figure 22. Losange.

- Variable et Path Modifier spécification :

Dans les règles, les deux premières classes **Path_Modifier** et **Variable** sont utilisées. **Path_Modifier** est une impasse. Si nous considérons l'enchaînement de liens causaux asymétriques et unidirectionnels comme un chemin, alors, ce chemin doit s'arrêter sur un **Path_Modifier** et ne pas démarrer depuis celui-ci. Autrement dit, il interrompt une relation

transitive. Sous la forme de règles, les relations ne peuvent démarrer ou ne traverser qu'une **Variable** et peuvent terminer sur une **Variable** ou un **Path_Modifier** (e.g., règle 1 et 2). Un **Path_Modifier** peut être relié avec une **Variable** via une relation symétrique bidirectionnelle telle que **Share_ancestor**. La classe **Path_Modifier** permet la représentation des relations non dirigées (règles (3) et (5)) et supprime les relations indirectes. (figure 23). Le domaine et le co-domaine (*range*), correspondant respectivement au concept de départ et d'arrivée, n'ont pas été utilisés, car il doit être possible pour un utilisateur de représenter une relation causale dirigée partant d'un Path_Modifier vers une Variable.

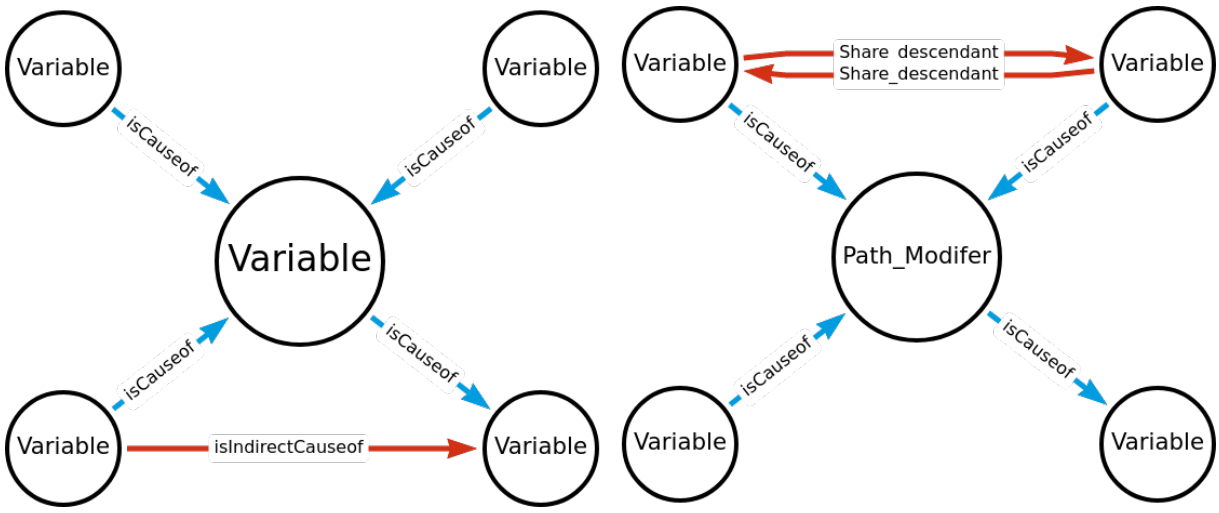


Figure 23. Share_descendant.

- Création d'une hiérarchie avec de nouvelles object properties pour un total de 14 (figure 24):

- *Related_to* est la plus haute des object properties de cette hiérarchie. Elle a pour synonyme corrélation statistique, ou association statistique. Ce qui signifie que, s'il existe un lien entre deux variables, alors, avec un nombre suffisamment important d'individu dans un échantillon, un lien statistiquement significatif sera retrouvé.

- *Directed_Relation* (→, → →, ↔, ↔→) englobe toutes les représentations avec une flèche dirigée (hormis les object properties inverses).

- *Indirect_Directed_Relation* (→ →, ↔, ↔ →) sont les relations causales dirigées qui relient deux variables en passant par une troisième.

- *Non_Directed_Relation* regroupe toutes les représentations non dirigées.

- *Inverse_Directed_Relation* regroupe les classes inverses au nombre de deux.

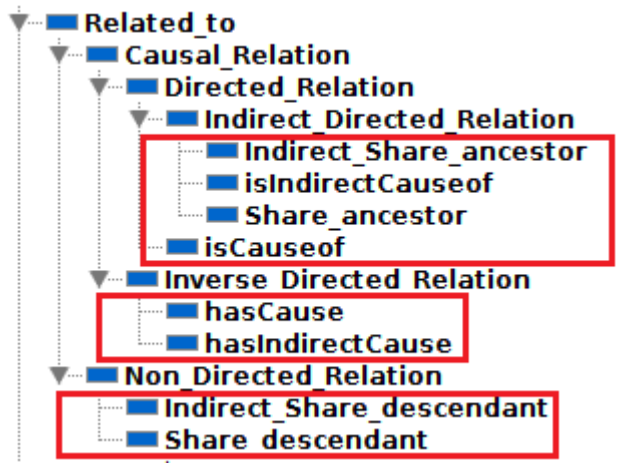


Figure 24. hiérarchie complète des object properties de OntoBioStat à ce stade de la construction. Sont encadrées en rouge les relations de base.

Cette nouvelle hiérarchie permet la construction de règles plus simples et d'avoir une granularité flexible sur une relation donnée entre deux variables. Concernant la simplification des règles, voici plusieurs exemples :

- La relation *IndirectCauseof(?x,?z)* est définie par trois situations : (isCauseof(?x,?y) ET isCauseof(?y,?z)) OU (isCauseof(?x,?y) ET isIndirectCauseof(?y,?z)) OU (isIndirectCauseof(?x,?y) ET isIndirectCauseof(?y,?z)). *IndirectCauseof(?x,?z)* peut être simplifiée par Inverse_Directed_Relation(?x,?y) ET Inverse_Directed_Relation(?y,?z) (règle (1)).

- La relation *Indirect_Share_descendant* est définie par plusieurs situations : une des relations est *isCauseof*, *isIndirectCauseof*, *IndirectShare_ancestor* ou une *Share_ancestor* (regroupés sous *Directed_Relation*) et l'autre relation peut être une *isIndirectCauseof*, *IndirectShare_ancestor* ou une *Share_ancestor* (regroupés sous *Indirect_Directed_Relation*) mais pas une *isCauseof* qui est exclue de *Indirect_Directed_Relation* (règle (5)).

Concernant la granularité des relations entre deux variables, les besoins varient en fonction de la question. La question peut être uniquement statistique, par exemple, « est-ce que deux variables sont corrélées ? » (*Related_to*) ou mécanistique, par exemple, « pourquoi elles sont reliées ? » (*isIndirectCauseof*).

Variable (?y) ∧ Variable(?x) ∧ Inverse_Directed_Relation(?y, ?x)

∧ Inverse_Directed_Relation(?z, ?y) → isIndirectCauseof(?x, ?z)

(1)

Variable(?z) ∧ Variable(?y) ∧ isCauseof(?z, ?x) ∧ isCauseof(?z, ?y) →

Share_ancestor(?x, ?y)

(2)

Path_Modifier(?z) ∧ Variable (?x) ∧ Variable(?y) ∧ isCauseof(?

y, ?z) ∧ isCauseof(?x, ?z) → Share_descendant(?x, ?y)

(3)

Variable(?y) ∧ Share_ancestor(?x, ?y) ∧ Inverse_Directed_Relation(?z, ?y) →

Indirect_Share_ancestor(?x, ?z)

(4)

Path_Modifier(?z) ∧ Variable(?y) ∧ Variable(?x) ∧

Directed_Relation(?y, ?z) ∧ Indirect_Directed_Relation(?x, ?z) →

Indirect_Share_descendant(?x, ?y) ∧ hasDescendant(?y, ?z) ∧ hasDescendant(?x, ?z)

(5)

Éviter les cycles

Pour éviter les cycles, c'est à dire lorsque l'exposition cause l'exposition ou la causalité inverse quand le critère de jugement cause l'exposition, les règles des relations causales ont été modifiées comme il suit :

Une relation causale ne peut pas passer par l'**Exposure** (e.g., si A→ **Exposure** → C alors l'exposition agit comme un Path_Modifier).

Une relation causale inférée ne ne peut pas démarrer depuis, ni passer par l'Outcome (e.g., Outcome → A → C alors l'Outcome ne cause pas indirectement C).

Il faut donc pouvoir manipuler l'Exposure et la Covariable en même temps. Elles sont regroupées sous la classe Explicative_Variable. Cette nouvelle classe (synonyme de variable indépendante) correspond aux variables qui se situent sur la droite de l'équation du modèle :

Outcome = Exposure + Covariable_1 + … + Covariable_n

Les règles (1), (2) et (4) sont modifiées.

Covariate(?y) ∧ Explicative_Variable(?x) ∧ Inverse_Directed_Relation(?y, ?x)

∧ Inverse_Directed_Relation(?z, ?y) → isIndirectCauseof(?x, ?z)

(1)

Covariate(?z) ∧ Variable(?y) ∧ isCauseof(?z, ?x) ∧ isCauseof(?z, ?y) →

Share_ancestor(?x, ?y)

(2)

Covariate(?y) ∧ Share_ancestor(?x, ?y) ∧ Inverse_Directed_Relation(?z, ?y) →

Indirect_Share_ancestor(?x, ?z)

(4)

Par ailleurs, il existe des cercles causaux comme par exemple dans la dépression avec aboulie entraînant culpabilité de ne rien faire entraînant une augmentation du niveau de dépression augmentant l'aboulie, etc. Dans ce cas, les utilisateurs des DAG représentent la variable aboulie à un instant t0 puis à un instant t0+1 ce qui en fait deux variables différentes (figure 25).

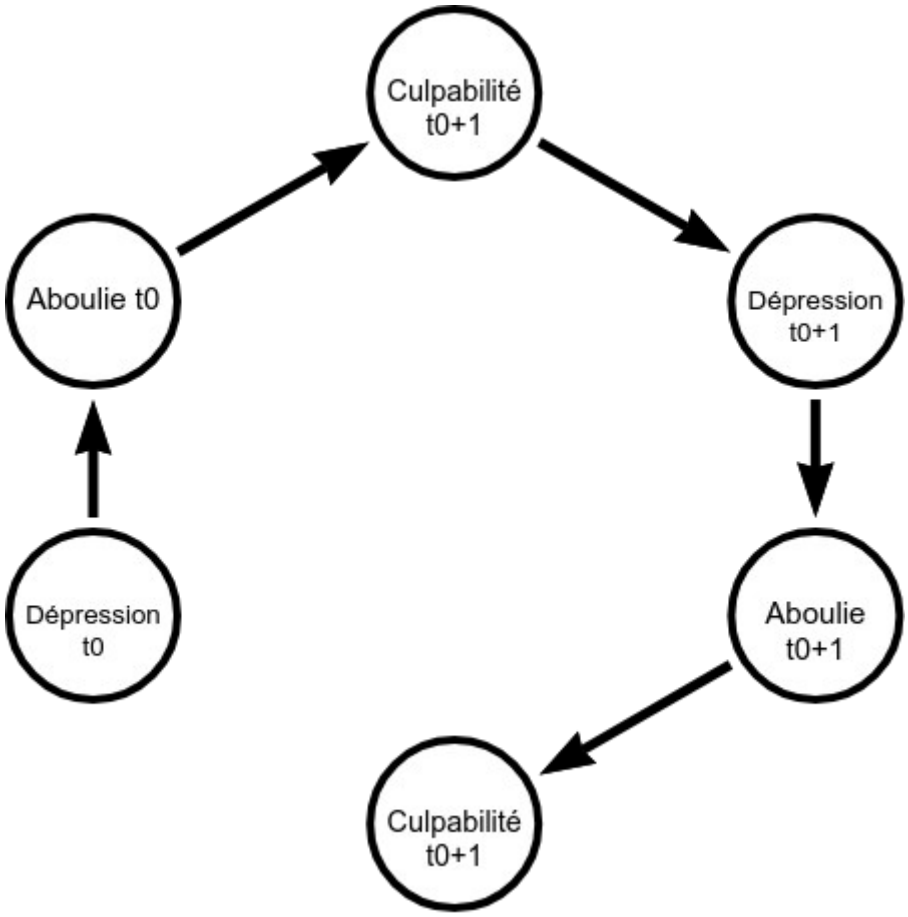


Figure 25. Exemple de cercle de causalité sous une forme de diagramme causal acyclique.

2.2.2. Classes inférées pour la sélection des variables

Variable de confusion, médiation et collision

La représentation sous la forme d'un diagramme causal permet de déduire si les variables sont des variables de confusion, de collision ou de médiation. Ces trois variables (**Confounder**, **Collider** et **Mediator**) ont été représentées en tant que sous classes de la classe **Covariate**. En effet, seules les covariables peuvent être des variables de confusion, collision ou médiation. La classification de celles-ci dépend des relations qu'elles entretiennent avec le couple **Outcome** et **Exposure**. De plus, ces trois classes ne sont pas disjointes car il est possible pour une instance d'appartenir à deux ou trois de ces classes. Par exemple, dans la figure 26, une variable de confusion est aussi une variable de collision.

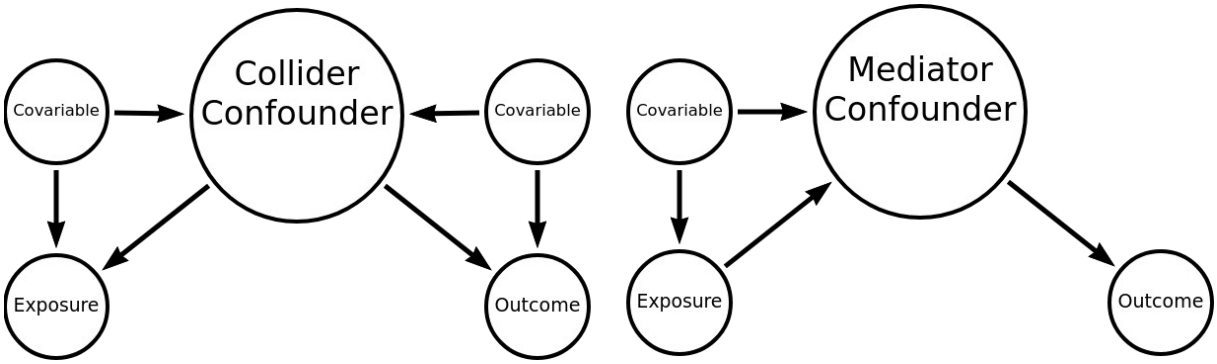


Figure 26. La première représentation est celle d'un collider qui est aussi un confounder papillon mediation.

Le corpus a permis de recueillir les synonymes de variable de médiation. Une variable de médiation (**Mediator**) est définie par l'axiome :

((hasCause min 1 Exposure_stressor) or (hasIndirectCause min 1 Exposure_stressor)) and ((isCauseof min 1 Outcome) or (isIndirectCauseof min 1 Outcome))

Une variable de collision (**Collider**) est définie par l'axiome :

Covariate and (hasCause min 2 Inferred_Variable)

Une variable de confusion peut être considérée comme une variable de confusion pour plusieurs raisons. Ainsi, plusieurs lignes de raisonnement ont été créées avec une classe correspondante pour chacune : **Direct_Confounder**, **Indirect_Confounder**, **Confounder_like**, **Collider_Confounder** (figure 27). Ces classes ne sont pas disjointes car il est possible pour une instance d'appartenir à plusieurs d'entre elles. Elles répondent toutes à la même question de compétence : « Quelles sont les variables qui confondent le vrai effet causal entre exposure et outcome ? »

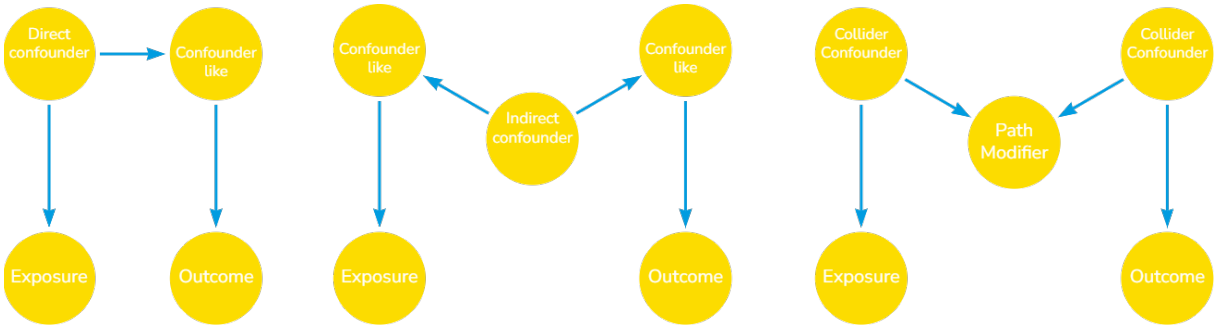


Figure 27. Classes de Confounder.

Direct_Confounder : Correspond aux variables de confusion directement en relation avec au moins l'Outcome, et/ou l'Exposure. Elle est définie par l'axiome suivant :

((isCauseof some Exposure_stressor) and (isCauseof some Outcome))

or ((isCauseof some Exposure_stressor) and (isIndirectCauseof some Outcome))

or ((isCauseof some Outcome) and (isIndirectCauseof some Exposure_stressor))

Indirect_Confounder : Correspond aux variables de confusion indirectement en relation avec l'Outcome et l'Exposure. Elle est définie par la règle SWRL suivante :

Inverse_Directed_Relation(?e, ?c) ^ Inverse_Directed_Relation(?o, ?b) ^ Covariate(?c) ^ Covariate(?b) ^ isCauseof(?a, ?b) ^ differentFrom(?c, ?b) ^ Exposure_stressor(?e) ^ differentFrom(?a, ?c) ^ Outcome(?o) ^ isCauseof(?a, ?c) -> Indirect_Confounder(?a)

La règle ne pouvait pas être : cause de manière indirect l'exposure et l'outcome. En effet, dans certains cas, cette règle aurait entraîné des inférences erronées. La véritable définition exclut toutes variables qui causent de manière indirecte l'exposition et l'outcome par l'intermédiaire du même descendant. Dans la figure ci contre, la covariable du haut n'est pas une variable de confusion (figure 28).

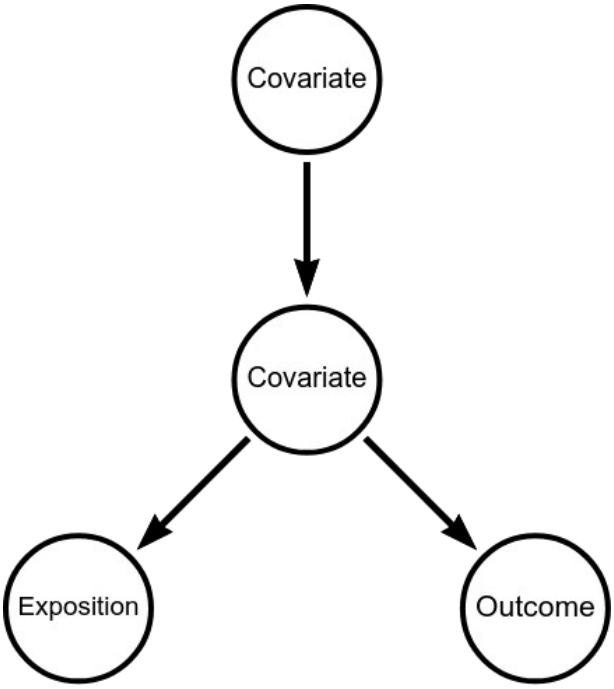


Figure 28. Exemple variable qui n'est pas une variable de confusion.

Confounder_like : Correspond à une variable sur le chemin causal entre une variable de confusion et l'Exposition ou l'Outcome. Elle est définie par l'axiome suivant :

Covariate **# C'est une covariate**

and (((Inverse_Directed_Relation some Direct_Confounder) or (Inverse_Directed_Relation some Indirect_Confounder) or (Inverse_Directed_Relation some Interaction_Confounder) or (isModifiedby some Interaction_Confounder)) **# Causée par une variable de confusion**

and ((isCauseof some Exposure_stressor) or (isCauseof some Outcome) or (isIndirectCauseof some Exposure_stressor) or (isIndirectCauseof some Outcome))) **# et qui cause l'exposition ou le critère de jugement**.

Collider_Confounder : Attention, le **Collider_Confounder** ne représente pas la situation où un collider est aussi un confounder. Il correspond à une variable qui devient une variable de confusion secondairement à l'ajustement, la stratification, l'exclusion d'une variable de collision. Elle est définie par la règle SWRL suivante :

hasDescendant(?b, ?d) ^ Covariate(?d) ^ Non_Directed_Relation(?a, ?c) ^ hasDescendant(?c, ?d) ^ Directed_Relation(?c, ?b) ^ Covariate(?c) ^ Exposure_stressor(?a) ^ Indirect_Share_descendant(?b, ?c) ^ hasDescendant(?a, ?d) ^ Outcome(?b) ^ differentFrom(?c, ?d) ^ Path_Modifier(?d) -> Collider_Confounder(?c)

La causalité inverse

La causalité inverse correspond à la situation où le critère de jugement cause de manière directe ou indirecte l'exposition. Dans de telles situations, en conservant les règles initiales, le raisonneur aurait renvoyé une erreur à cause de l'anti symétrie qui aurait été violée (i.e., A cause B qui cause C qui cause A). Après modification des règles, il ne renvoie pas d'erreur et il n'est plus possible de repérer ce problème (i.e., les inférences considèrent l'outcome comme un path modifier, les cycles sont impossibles). La création d'une classe spécifique avec une règle dédiée s'est donc imposée :

Pour tenir compte d'une possible causalité inverse existante et être capable de la repérer la classe **Reverse_Causality** sous classe de **Outcome** a été créée**.** Elle est inférée grâce à la règle SWRL suivante :

Inverse_Directed_Relation(?c, ?o) ^Inverse_Directed_Relation(?e, ?c) Covariate(?c)^Exposure_stressor(?e) ^ Outcome(?o) -> Reverse_Causality(?o)

Par exemple, lorsque le lien entre antidiabétiques oraux (exposition) et cancer du pancréas (critère de jugement) est recherché, il se peut que ce soit le cancer du pancréas qui soit à l'origine d'une dysfonction pancréatique ayant entraîné une prescription d'anti diabétique oral.

Variables de confusion non mesurées et variables proxy (ou intermédiaires)

Les variables de confusion non mesurées entraînent l'impossibilité de corriger les biais. D'autres variables sur le chemin causal tel que les Confounder-like définis plus haut peuvent suffire à corriger les biais. Cependant, il arrive que l'ajustement sur la variable de confusion non mesurée soit la seule option. Les variables « proxy » ou variables intermédiaires sont des causes (proxy ascendant), des conséquences (proxy descendant) ou partagent le même ancêtre que la variable de confusion non mesurée. On peut simplement dire que c'est une variable fortement corrélée (Related_to). En utilisant ces variables intermédiaires à la place des variables de confusion non mesurées, il est possible de corriger partiellement voire complètement le biais. L'identification des variables non mesurées est faite grâce au data-type property isMeasured (TRUE/FALSE). L'identification des variables intermédiaires est faite grâce à l'existence d'un object property entre celle-ci et la variable non mesurée. De plus, la variable intermédiaire ne doit pas causer l'exposition et le critère de jugement. L'object property NotCauseof est créée dans ce but, afin de rendre explicite l'absence de relation causale entre deux variables (figure 29). Le terme anglais proxy

continuera d'être utilisé car un des synonymes de *médiator* en anglais est variable intermédiaire.

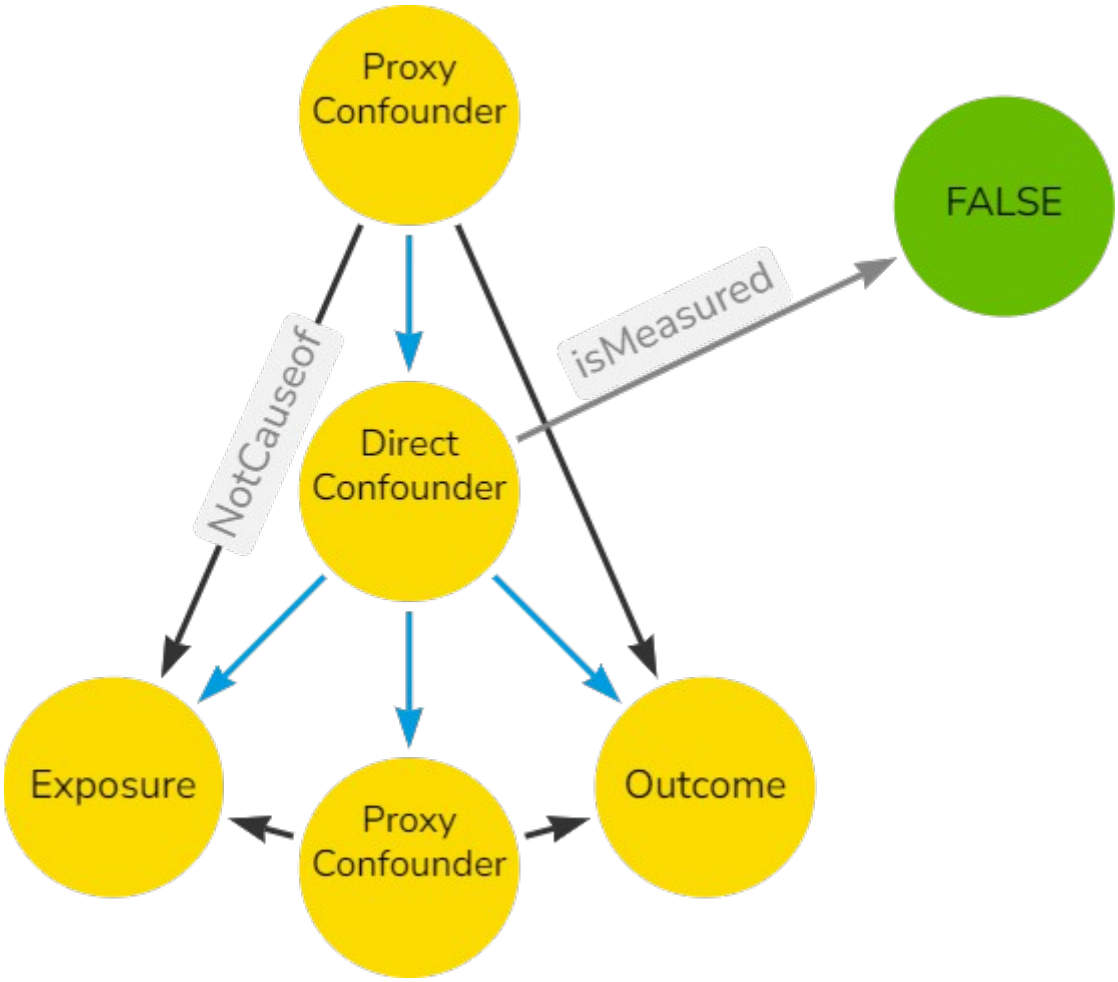


Figure 29. Proxy_Confounder, flèches noires: NotCauseof, flèches bleues: isCauseof.

Elles sont inférées grâce à la règle SWRL suivante :

NotCauseof(?b, ?e) ^ Inverse_Directed_Relation(?e, ?a) ^ isMeasured(?a, false) ^ Inverse_Directed_Relation(?o, ?a) ^ Covariate(?b) ^ Covariate(?a) ^ NotCauseof(?b, ?o) ^ isCauseof(?a, ?b) ^ Exposure_stressor(?e) ^ Outcome(?o) -> Proxy_Confounder(?b)

Question de compétence : Existe-t-il des variables proxies de variable de confusion ?

Figure 30: Synthèse des classes du cadre de construction minimal

2.2.3. Ajout de relations et inférences correspondantes

La recherche bibliographique sur le sujet des CD et des DAG a permis d'extraire des relations qui n'ont pas été présentées dans l'introduction sur les DAG (interaction ou des spécifications de relation (relation causale positive ou négative)). Chaque paragraphe présente la ou les définitions ainsi que les représentations graphiques retrouvées dans les articles avant de les intégrer dans l'ontologie.

Interactions et relations

L'inclusion ou non d'une interaction (aussi appelée effect modification) dans le modèle statistique fait partie intégrante du processus de sélection des variables. Les interactions sont souvent modérées ce qui entraîne un faible usage de celles-ci. Cependant, en tant que partie intégrante du processus de la sélection des variables, il est indispensable de pouvoir proposer sa représentation ontologique. Différents articles sur les graphes orientés acycliques ont fourni des représentations d'interaction différentes (VanderWeele TJ , 2007; Weinberg CR, 2007 ; Nilsson A et al., 2021; Lopez PM et al., 2019). Après avoir fait le point sur ces représentations existantes, une forme ontologique sera proposée.

Weinberg définit quatre situations (figures 31-34) avec graphiques et exemples à l'appui, dans lesquelles les interactions peuvent être impliquées : (i) Pure effect modification : A et B ont un effet sur C si et seulement si ils sont présents tous les deux ; (ii) soit A une variable indépendante de B et C qui interagit sur la relation entre B et C ; (iii) soit A une variable qui cause C et interagit avec la relation BC, B une variable qui cause C et interagit avec la relation AC ; (iv) soit A une variable qui cause B et interagit avec la relation BC et B une variable qui cause C.

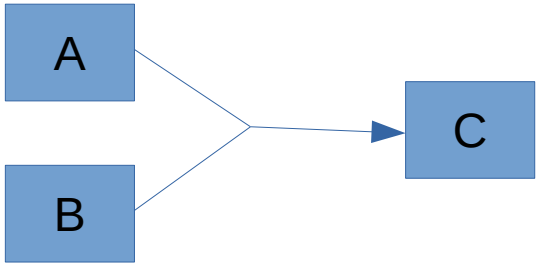


Figure 31: Situation 1 : Pure effect modification

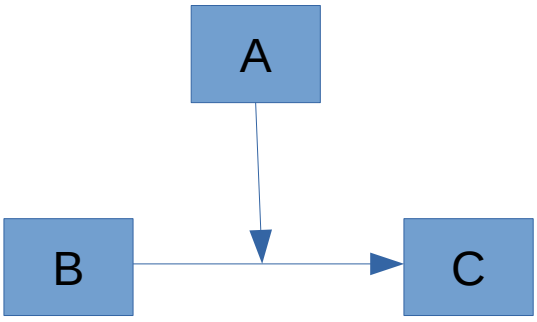


Figure 32: Interaction situation 2

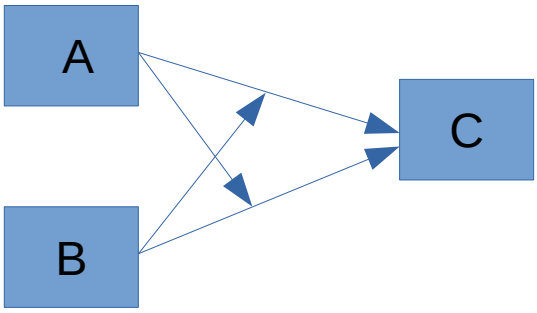


Figure 34: Interaction situation 3

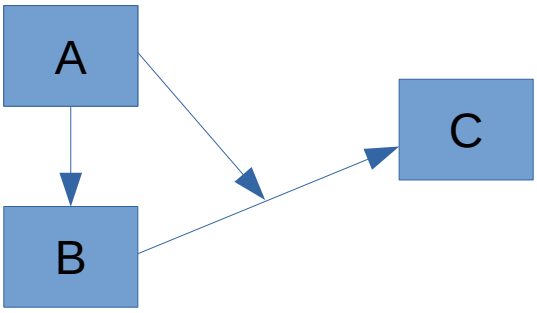


Figure 33: Interaction situation 4

La situation 2 correspond par exemple à l'effet du phénotype foncé par rapport au clair (i.e ;personnes à la peau clair et aux cheveux roux) sur la probabilité d'avoir un cancer de la peau en cas d'exposition solaire. Le phénotype en lui même sans le soleil

n'est pas (ou peu) responsable de cancer de la peau (fig 32). La troisième situation représente par exemple l'effet de l'alcool et du tabac indépendamment l'un de l'autre et l'effet synergique de ceux-ci sur le risque de survenue d'un cancer de la sphère ORL (fig 34).

La situation 4 est celle qui est la plus importante pour la sélection des variables, car A se place en tant que variable de confusion. La situation 2 permet de connecter deux parties d'un graphe.

VanderWeele et al. décrivent 4 types d'interaction (figure 35) : (i) direct effect modifiers; (ii) indirect effect modifiers; (iii) effect modifiers by proxy; et (iv) effect modifiers by common cause. Ces définitions d'interactions sont des descriptions purement graphiques et ne reposent pas sur un rationnel biologique contrairement à Weinberg. Ci-dessous les représentations directement issues de l'article. Pour chaque figure, le type d'interaction défini correspond à la variable encadrée. Par exemple, dans la figure 1 le direct effect modification est le X. Il est spécifié que X est une interaction entre E et D alors que cette information n'est pas représentée. La représentation graphique de l'interaction n'est pas différente de celle d'une simple relation causale.

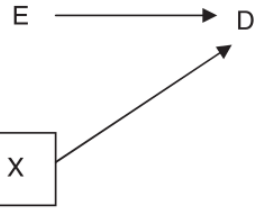


FIGURE 1. Direct effect modification: E, drug exposure; D, hypertension outcome; X, genotype, a direct effect modifier.

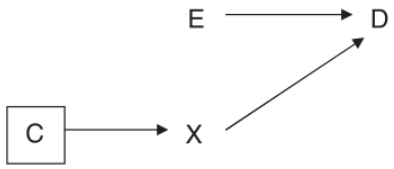


FIGURE 2. Indirect effect modification: E, drug exposure; D, hypertension outcome; X, genotype; C, mother's genotype, an indirect effect modifier.

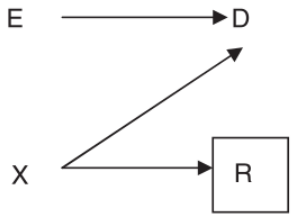


FIGURE 3. Effect modification by proxy: E, drug exposure; D, hypertension outcome; X, genotype; R, hair color, an effect modifier by proxy.

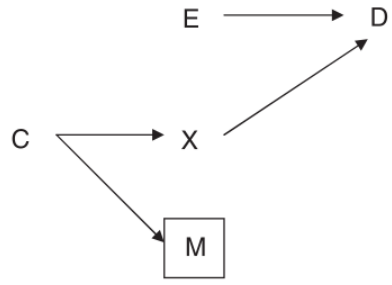


FIGURE 4. Effect modification by common cause: E, drug exposure; D, hypertension outcome; X, genotype; C, mother's genotype; M, mother's hair color, an effect modifier by common cause.

Figure 35: Compilation des figures de l'article VanderWeele et al., 2007

Nilson et al. proposent une représentation appelée Interaction DAG (IDAG) et définissent aussi quatre concepts : (i) total interaction, (ii) direct interaction, (iii) indirect interaction, et (iv) confounded interaction. En ce qui concerne les trois premiers, leur définition est transposable à une situation de médiation si ce n'est que dans un IDAG la

relation sur laquelle les variables interagissent correspond à un seul nœud représenté avec un delta et le nom des deux nœuds. Par exemple, si on s'intéresse à l'effet de A sur Y et qu'il existe une variable X qui intéragit sur la relation entre A et Y alors X → deltaYA. X est une interaction directe et dans la situation X → Z → delta YA, X est une interaction indirecte. Le dernier concept aussi appelé effect modification by proxy correspond à l'effect modification by common cause de VanderWeele. Q ← X → deltaYA.

Lopez PM et al. redéfinissent l'interaction comme un mécanisme de médiation. Ainsi, la situation 2 de Weinberg peut être représentée avec un nœud intermédiaire (figure 36).

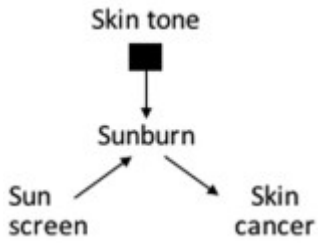


Figure 36. Figure extraite de l'article Lopez PM et al., 2019

Les nouvelles classes inférées (figure 37) et les object properties de OntoBioStat :

La création des classes et des object properties a été guidée par les exemples et les graphiques de l'article de Weinberg. Dans OntoBioStat, une interaction est considérée comme une instance qui interagit avec deux variables (Interact_with exactly 2 Inferred_Variable). Dans OntoBioStat, les relations ne sont pas des classes mais des objects properties. Il n'est donc pas possible formellement de dire : 'une variable interagit avec une relation'. Les interactions peuvent impliquer un couple de variables qui agissent sur une autre comme le **Pure_effect_modification** ou une seule variable **Interaction_Single**. Les object properties *Interact_with* et son inverse *isModifiedby* sont créées dans le but d'inférer tous les sous types d'interaction ainsi que **Interaction_Confounder** qui correspond à une interaction qu'il faut inclure dans le modèle. La classe **Synergistically_Antagonistically** correspond à la situation 3. Elle contient les instances qui peuvent soit interagir en potentialisant l'effet, soit en antagonisant l'effet tout en ayant un effet causal propre sur le critère de jugement (figure 38).

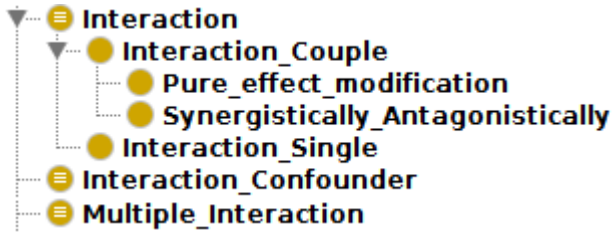


Figure 37: Classes d'interaction

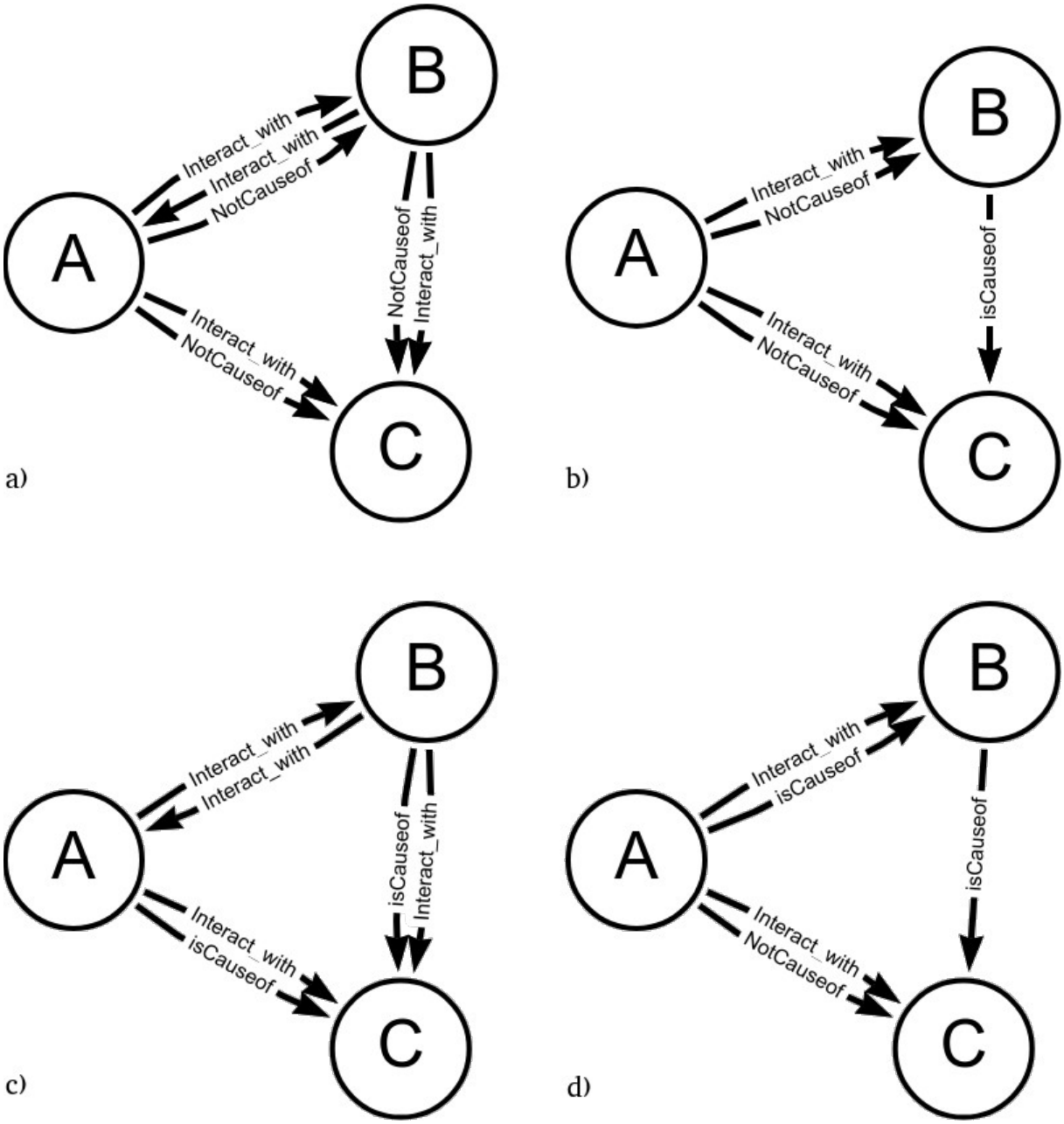


Figure 38. Les différentes types d'interaction représentés.

Les règles SWRL correspondantes sont les suivantes (figure 38) :

Pure effect modification (Situation 1, a) :

NotCauseof(?a, ?c) ^ NotCauseof(?a, ?b) ^ Interact_with(?b, ?a) ^ NotCauseof(?c, ?b) ^ Interact_with(?a, ?b) ^ Interact_with(?b, ?c) ^ Interact_with(?a, ?c) ^ differentFrom(?c, ?b) ^

differentFrom(?a, ?b) ^ differentFrom(?a, ?c) -> Pure_effect_modification(?a) ^ Pure_effect_modification(?b)

Interaction Single (Situation 2, b) :

NotCauseof(?a, ?b) ^ NotCauseof(?a, ?c) ^ Interact_with(?a, ?b) ^ Interact_with(?a, ?c) ^ differentFrom(?b, ?c) ^ differentFrom(?a, ?b) ^ differentFrom(?a, ?c) ^ isCauseof(?b, ?c) -> Interaction_Single(?a)

Synergistically_Antagonistically (Situation 3, c) :

Interact_with(?b, ?a) ^ Interact_with(?a, ?b) ^ Interact_with(?b, ?c) ^ Interact_with(?a, ?c) ^ isCauseof(?b, ?c) ^ differentFrom(?b, ?c) ^ differentFrom(?a, ?b) ^ differentFrom(?a, ?c) ^ isCauseof(?a, ?c) -> Synergistically_Antagonistically(?b) ^ Synergistically_Antagonistically(?a)

Interaction_Confounder (Situation 4, d) :

Indirect_Directed_Relation(?b, ?o) ^ Interact_with(?a, ?b) ^ Interaction(?a) ^ Directed_Relation(?a, ?e) ^ Exposure_stressor(?e) ^ Outcome(?o) -> Interaction_Confounder(?a)

Ces règles permettent de répondre à la question de compétence « Existe-t-il une interaction qui pourrait biaiser le vrai effet causal entre exposure et outcome ? ».

C'est une représentation qui reste lourde en terme de spécification et de relation pour un ontologie d'application. Une représentation alternative pourrait impliquer des instances d'interaction correspondant à l'effet causal joint de deux variables. En tant qu'ontologie d'application, c'est l'usage qui pourra aider à trancher entre les deux représentations. Comme, les situations qui impliquent l'utilisation d'interactions sont moins courantes que celles qui impliquent des variables de confusion classiques, il se peut que la représentation première reste inchangée.

Les relations signées

En épidémiologie, pouvoir anticiper le sens du biais ou connaître le sens du biais lorsqu'il reste des variables de confusion sur lesquelles nous n'avons pas pu ajuster (parce que non mesurées) est un point important. Pour se faire, des représentations de signes négatifs ou positifs ont été ajouté sur les graphes orientés acycliques (VanderWeele TJ et al., 2008 ; VanderWeele TJ et al., 2010 ; VanderWeele TJ et al., 2012).

Les DAG signés correspondent à des DAG pour lesquels ont connaît le signe de la relation : « positif » signifie que le risque augmente et « négatif » signifie que le risque diminue. En d'autres termes, indiquer le signe d'un lien causal (*signed edge*), c'est définir si la présence d'un traitement, d'une exposition ou autre va diminuer ou augmenter l'occurrence d'un critère de jugement.

VanderWeele TJ définit quatre concepts : « monotonic effect », « average monotonic effect », « weak monotonic effect » et « signed edge ». Un *Monotonic effect* positif (ou négatif) correspond à un effet strictement positif ou neutre (ou strictement négatif ou neutre) au niveau **individuel** alors que le *average monotonic effect* correspond à la version **populationnelle** du *monotonic effect*. C'est à dire qu'en cas de positive *average monotonic effect*, l'estimation de l'effet causal (rapport des côtes ou risque relatif) sera toujours positif. On parle aussi d'effet moyen, dans le sens où individuellement certaines personnes auront un effet négatif, nul ou positif mais, en moyenne, l'effet sera positif. Par exemple, pour certains patients le traitement A aura un effet délétère alors que pour d'autres, il améliorera leur pronostic et en moyenne il sera meilleur que le B. Alors que le tabac sur la survenue d'un cancer du poumon a, au niveau individuel, soit un effet nul, soit il entraîne le cancer du poumon. Le *weak monotonic effect* positif correspond à un effet quasi-strictement positif ou neutre au niveau **individuel**.

Les relations signées dans OntoBioStat : les relations signées regroupent *Increase*, *Decrease*, *Contraindication* et *Absolute_indication*.

Increase et *Decrease* correspondent respectivement au positive *average effect* et negative *average effect*. Pour des raisons de simplification, les monotonic effect ou weak monotonic effect n'ont pas été représentés. Puisque un monotonic effect ou weak monotonic positif, implique un *average effect* positif, il est possible de les regrouper sous un même terme. En effet, l'information importante est est-ce que 'en moyenne' une variable donnée augmente ou diminue le risque. *Increase* et *Decrease* permettent, avec des règles adaptées, de déduire le sens du biais qu'entraîne une variable de confusion, et répondent donc à la cinquième question de compétence : "Quelle est la direction du biais causée par un facteur de confusion?".

Exemple 1 : l'obésité est la variable d'exposition d'intérêt et la nutrition la variable de confusion. Un mauvais comportement alimentaire (nutrition=1) va causer + d'obésité et + d'évènement cardiovasculaire. Nous aurons intérêt à ajuster sous peine de surestimer (+*+) l'effet de l'obésité (fig 39).

NB : Si nous modifions la catégorie de référence de la variable nutrition (bon comportement alimentaire <=> 1) elle diminue le risque d'obésité (-) et d'évènement cardiovasculaire (-). Nous aurons intérêt à ajuster sous peine de surestimer (-*-) l'effet de l'obésité.

Les neuf règles SWRL présentées ci-dessous incluent des differentFrom pour toutes les variables ainsi que les mêmes spécifications que les règles pour la causalité avec **Covariate**, **Variable** et **Path_Modifier**. Pour éviter une surcharge visuelle une version simplifiée des règles est présentée.

Decrease(?b, ?c) ^ Increase(?a, ?b) -> Decrease(?a, ?c)

Increase(?a, ?b) ^ Increase(?b, ?c) -> Increase(?a, ?c)

Decrease(?a, ?b) ^ Decrease(?b, ?c) -> Increase(?a, ?c)

Exemple 2 : Le signe de l'effet de l'exposition médié par une variable de intermédiare est inféré de la même manière que pour la confusion (fig 39). Par exemple, si un mauvais comportement alimentaire va augmenter la probabilité de devenir obèse et que le fait d'être obèse augmente la probabilité de l'occurrence d'un évènement cardiovasculaire alors l'effet médié est positif.

La différence avec les règles SWRL précédentes c'est qu'il existe deux conséquent au lieu d'un car Share_ancestor est symétrique.

Decrease(?a, ?c) ^ Decrease(?a, ?b) -> Increase(?c, ?b) ^ Increase(?b, ?c)

Increase(?a, ?c) ^ Increase(?a, ?b) -> Increase(?c, ?b) ^ Increase(?b, ?c)

Decrease(?a, ?c) ^ Increase(?a, ?b) -> Decrease(?b, ?c) ^ Decrease(?c, ?b)

Exemple 3 : Le signe de l'effet entre deux variables qui partagent un descendant commun est négatif si le lien causal entre ces variables et le descendant était positif. Par exemple, concernant la taille d'une tumeur bénigne et la symptomatologie d'une tumeur qui pousse à l'opération, parmi les personnes opérées le lien entre taille et symptomatologie sera négatif (fig 39).

La différence avec les trois règles SWRL précédentes, c'est l'inversion du signe à cause du Path_Modifier.

Decrease(?c, ?a) ^ Decrease(?b, ?a) ^ -> Decrease(?c, ?b) ^ Decrease(?b, ?c)

Increase(?c, ?a) ^ Increase(?b, ?a) -> Decrease(?c, ?b) ^ Decrease(?b, ?c)

Decrease(?c, ?a) ^ Increase(?b, ?a) -> Increase(?c, ?b) ^ Increase(?b, ?c)

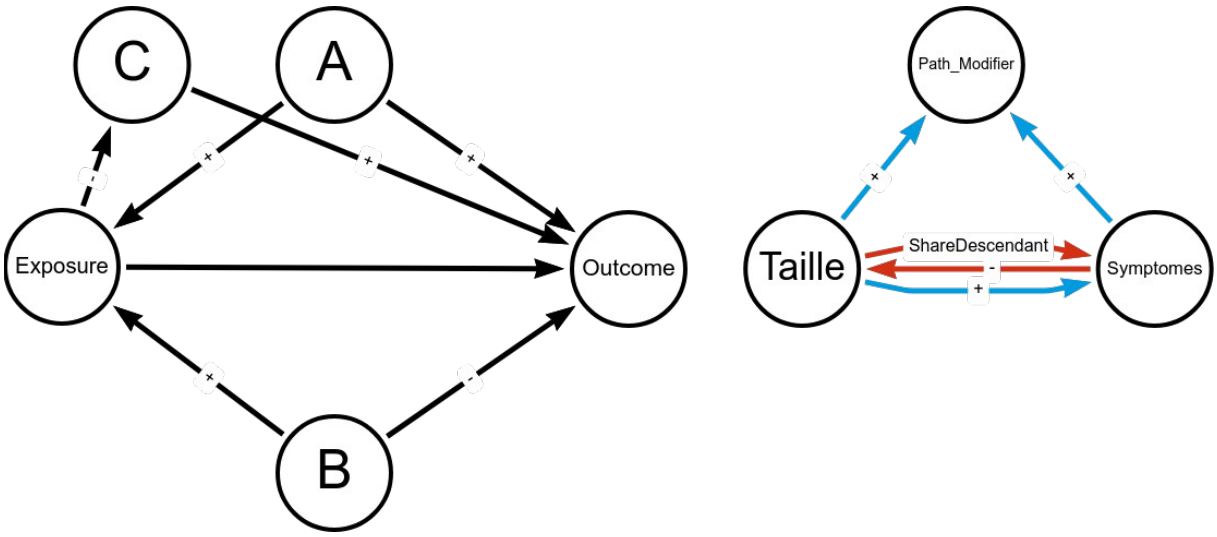


Figure 39 Diagrammes causaux avec relations signées.

Utilisations : Pour un **Confounder** donné, l'utilisateur d'OntoBioStat regarde le signe de sa relation avec le critère de jugement et l'exposition si ++ ou - - alors, il biaise vers le haut si - + ou + - alors il biaise vers le bas. L'intérêt est essentiellement de fournir à l'utilisateur une idée du sens du biais quand il ne peut pas corriger le facteur de confusion.

Limites : Lorsqu'il existe déjà un lien causal entre taille et symptomatologie (+) et qu'un nouveau lien causal *Share_descendant* (-) est inféré, il n'est pas possible de savoir si le lien entre taille et symptomatologie reste positif, devient négatif ou nul (fig 39). De même, si une variable de confusion confond l'effet causal entre exposition et critère de jugement via différentes inférences, celles-ci peuvent impliquer des relations causales avec des signes différents et il ne sera pas possible d'inférer le signe du biais. Les inférences se limitent donc à définir le signe de la relation entre variables et pas si le biais est vers le haut, le bas ou nul.

Contraindication et *Absolute_indication* sont des causes dites suffisantes et (quasi-)déterministes. La définition d'une cause suffisante d'après le MeSH est « cause are sufficient when they initiate or produce an effect ». En cas de présence d'un facteur de contre indication, le traitement A ne sera administré à aucun patient mais, en cas d'absence de facteur de contre indication, le traitement A aussi bien que le traitement B pourront être administrés. Ce n'est pas la définition du *monotonic effect*. En effet, cette définition correspond à l'exemple suivant : le tabac au niveau individuel ne fait qu'augmenter le risque de cancer du poumon et pour aucun des fumeurs l'arrêt de la cigarette n'entraînera un cancer du poumon. Dans le cadre de la médecine elles sont rares, sauf dans les protocoles de soin. Ce qui explique pourquoi, ces deux objects properties ont un nom en lien avec les indications.

Lorsqu'une variable est une contre indication et qu'elle est aussi impliquée dans une relation causale avec le critère de jugement, alors, il ne faut pas ajuster sur celle-ci, mais bien retirer les patients du jeu de données. Ces objects properties permettent d'inférer avec une règle SWRL une nouvelle classe **Unadjusted_Confounder** qui répond à la question : « Faut-il exclure des patients qui ne seraient pas comparable aux autres ? »

Cette nouvelle classe est définie par l'axiome suivant :

(((Absolute_Indication some Exposure_stressor) or (Contraindication some Exposure_stressor)) #La variable est une contre indication ou une indication absolue de l'exposition

and (Causal_Relation some Outcome) #Elle a une relation causale avec le critère de jugement

and (isCauseof some Exposure_stressor)) #La variable est bien une cause de l'exposition

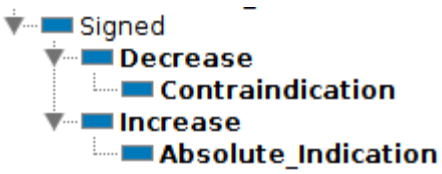


Figure 40: Nouvelles relations signées

Données manquantes, sélection des variables et relation causales

La ***gestion des données manquantes,*** qui prend en compte les mécanismes causaux qui ont généré des valeurs manquantes, fait partie des tâches à réaliser en amont de la sélection des variables. En effet, en cas d'analyse en cas complet, il suffit d'une valeur manquante sur une des variables d'ajustement pour que le sujet ou l'observation ne soit pas pris en compte dans le modèle (*leastwise* ou *complete case analysis*). La sélection d'une variable avec un pourcentage de données manquantes entraîne donc une perte de puissance statistique. De plus, la génération des données manquantes n'est parfois pas aléatoire, ce qui peut entraîner des biais dans l'estimation de l'effet de l'exposition (Westreich D et al., 2012).

Rubin DB 1976 a défini trois types de mécanismes de génération des données manquantes. Manquant complètement au hasard (MCAR), manquant au hasard (MAR) et manquant non du au hasard (MNAR). L'intitulé est trompeur, il faut donc se reposer sur les définitions suivantes. Lorsque les données manquantes sont complètement dues au hasard (MCAR) c'est qu'il est impossible de prédire leur présence quel que soit l'information dont on dispose. La raison est véritablement le hasard. Par exemple, une machine d'analyse sanguine fait une erreur dans un cas sur cinquante. Lorsque les données manquantes sont dues au hasard (MAR) c'est qu'il est possible de prédire leur présence avec les autres variables dont on dispose. Par exemple, si le sang du patient est acide, la mesure d'un marqueur

protéique X sera plus souvent manquante. Lorsque les données manquantes ne sont pas du hasard (MNAR) c'est que la donnée manquante est causée par la vraie valeur de la variable. Par exemple, quand le marqueur protéique X est trop élevé la machine a tendance à renvoyer une erreur.

Outre leur utilité dans la détection de variables de confusion et dans la protection contre les variables de médiation ou de collision, les graphes peuvent être utilisés dans la gestion des données manquantes.

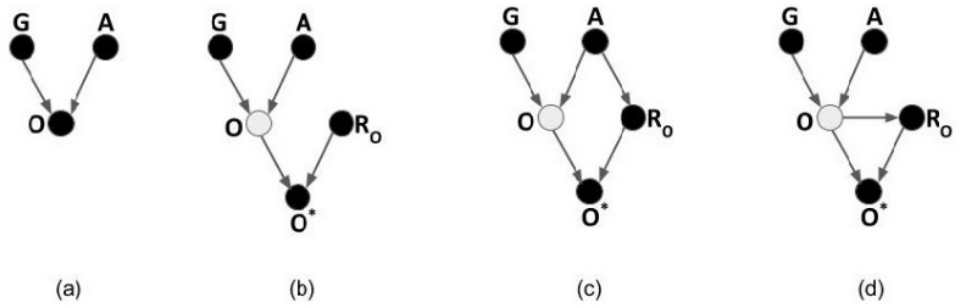


Figure 1: (a)causal graph under no missingness (b), (c) & (d) m-graphs modeling distinct missingness processes.

Figure 41: M-graph de Mohan K et al., 2021

Dans cette exemple (figure 41) issu de Mohan K et al., 2021, il y a quatre graphes qui représentent quatre situations de données manquantes. Dans la figure (a) la variable G et A causent la variable O et il n'y a pas de donnée manquante. Dans les figures suivantes la variable O* (variable intermédiaire de O aux valeurs incomplètes) est causée par O et R (variable qui définit la présence ou l'absence de la variable O, missingness mechanism). Dans la (b) R n'est causée par personne ce qui sous-tend l'idée que le mécanisme qui a mené à une valeur manquante ne dépend d'aucune caractéristique et donc est complètement aléatoire (MCAR). Dans la figure (c) et (d) on retrouve respectivement les situations MAR et MNAR.

La classe **Missing value reasons** est créée. Elle représente la raison de la présence de données manquantes pour une variable donnée. Une instance du nom de la variable concernée est créée en accolant le terme « NA » qui est souvent utilisé pour spécifier l'existence d'une valeur manquante. Cette instance est ensuite classée dans **Missing_value_reasons**. La relation *isCauseofNA* est créée. Elle a pour co-domaine (range) les instances de la classe **Missing_value_reasons** et pour domaine une variable explicative ou le critère de jugement

(**Outcome**). Elle permet d'inférer trois nouvelles classes (**MNAR**, **MCAR**, et **MAR**) qui sont des sous-classes de **Inferred_Missing_data_mechanism.**

Soit A la variable avec données manquantes, A_NA l'instance de Missing_value_reasons, X une autre variable. Si A cause A_NA alors A_NA est classée dans MNAR. Si X cause A_NA alors A_NA est classée dans MAR (figure 42). MAR et MNAR sont disjointes avec MCAR mais pas entre elles. En effet, il est possible que la raison de la présence de données manquantes soit à la fois MAR et MNAR.

Les règles SWRL permettant ces inférences sont :

Missing_value_reasons(?x)^Covariate(?y)^isMissing_Value_Reasonsof(?x,?y)^isCauseofNA(?y,?x) → MNAR(?x)

Missing_value_reasons(?x)^Covariate(?y)^Covariate(?z)^isMissing_Value_Reasonsof(?x,?y)^isCauseofNA(?z,?x) → MAR(?x)

z doit être différent de y.

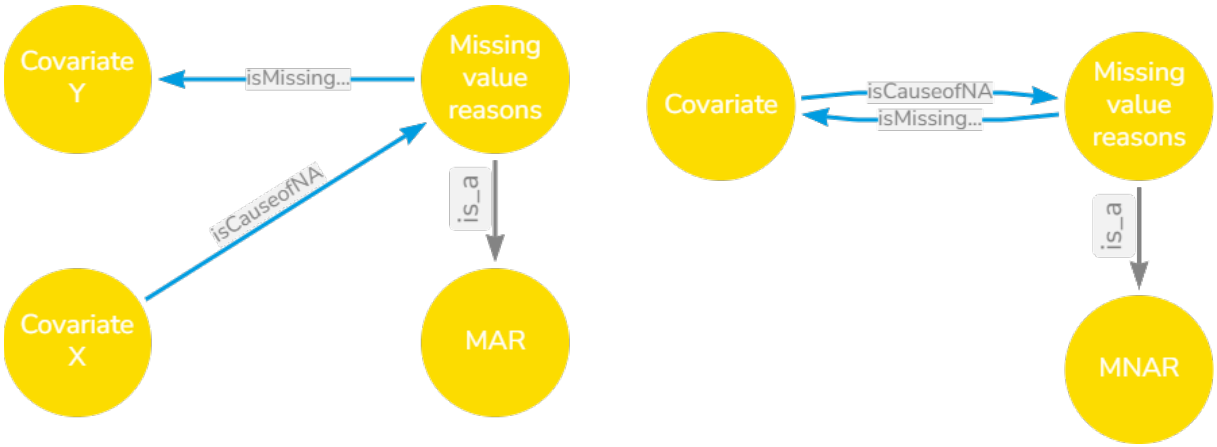


Figure 42: Exemples impliquant les classes et relations pour l'inférence des mécanismes de génération des données manquantes

Question de compétence : « Est-ce que le mécanisme responsable de la présence de données manquantes pourrait biaiser l'estimation du vrai effet causal entre exposure et outcome ? »

Dans cette représentation il existe deux nœuds au lieu de trois dans Mohan K et al., 2021 car il aurait fallu rajouter un troisième nœud qui aurait alourdi la représentation d'OntoBioStat sans aucun gain particulier au niveau inférence. La relation causale isCauseofNA permet d'éviter la prise en compte de la relation causale Covariable → Missing_value_reason dans les inférences pour les Confounders.

Exemple récapitulatif :

Soit la taille de la nécrose cérébrale une variable qui a des valeurs manquantes dépendant de sa propre valeur. En effet, on peut imaginer que quelqu'un qui aura une trop grosse nécrose

sera soit dans un état trop grave pour avoir une IRM (pas de mesure) soit aura une IRM avec une taille difficile à mesurer (mesure imprécise ou impossible de mesurer). Nous avons donc la Covariate taille de la nécrose qui a une Missing_value_reasons taille de la nécrose_NA et la taille de la nécrose qui est responsable de la donnée manquante (isCauseofNA) taille de la nécrose_NA. Taille de la nécrose est donc MNAR.

2.2.4. Un cadre de construction enrichi

Afin d'aider les biostatisticiens dans la construction du diagramme, il fallait représenter toute la connaissance nécessaire et potentiellement disponible pour la sélection des variables. Cette connaissance devait être générique, c'est à dire applicable à n'importe quelle étude ou jeu de données dont le but est de répondre à une question de recherche causale. Les classes, relations et les data-object properties décrites ci-après sont censées fournir un canevas pour la construction systématique et formelle d'un diagramme causal. De plus, à l'heure des données de grandes dimensions survient le problème de « data-fusion » (Bareinboim E et al., 2016). Ces représentations ontologiques permettent de tenir compte de l'hétérogénéité des jeux de données qui pourraient être utilisés comme un seul. Les nouvelles classes abordées dans cette partie sont intégrées dans la hiérarchie (figure 43).

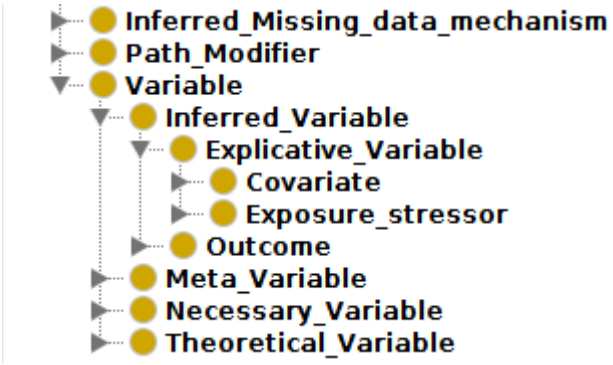


Figure 43: Hierarchie du cadre minimal avec les classes du cadre enrichi

Les méta-variables

La classe des **Meta_Variable** a été créée pour regrouper les connaissances dont on peut disposer pour une **Variable.** Ce sont des variables qui permettent de tenir compte du contexte d'une étude et enrichir la représentation d'une variable avec des informations complémentaires liées par exemple à la méthode de mesure ou le point de vue du transcripteur de la donnée. Les instances peuvent elles-mêmes être des variables, être causées, ou causer d'autres instances de **Variable**. Elles peuvent donc être impliquées dans la confusion du chemin causal entre **Exposure** et **Outcome**. Pour chacune d'entre elles, un object property a

été créé (e.g., la classe **Period** est reliée à une variable par l'object property *hasPeriod*). L'object property qui rassemble tous les object properties des métavariables est isMetavariableof. Ceux-ci ont été placés en tant que fils de *isCauseof* (Figure 44).

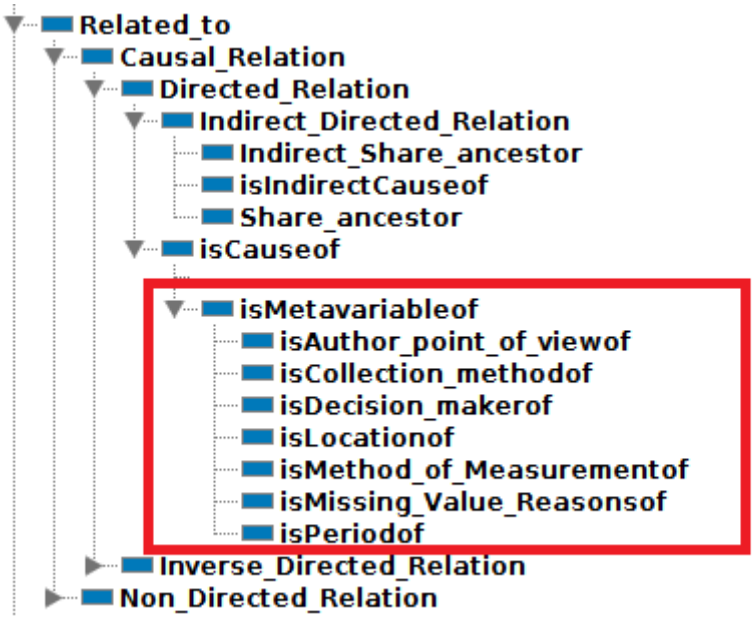


Figure 44: Nouvelles relations des Méta Variable intégrées dans les relations existantes

Si ces variables ne sont pas constantes alors elles sont considérées comme des covariables. Si elles sont constantes alors, elles sont considérées comme un Path_Modifier. Les règles SWRL suivantes ont été créées :

isMetavariableof(?x,?y)^hasPossibleValue(?x,'Complete')→ Covariate(?x)

Concernant ces méta-variables, il faut être attentif à deux éléments en particulier :

Ces variables, comme les variables d'indication de traitement, sont celles qui ont le plus de risque d'être des causes suffisantes. Par exemple, chaque fois que le chirurgien Z voit un patient Y, il décide toujours de l'exposer à X. Dans ce cas là, il faudra spécifier que le lien entre l'instance chirurgien de decision-maker et l'exposition X correspond à *ContraIndication* ou *AbsoluteIndication*.

Ces variables sont impliquées dans des phénomènes de causalité inverse. Par exemple, dans une étude cas-témoins, la méthode de collection pourrait être grandement influencée par la connaissance du statut malade ou non malade. En effet, la connaissance du statut va pousser l'investigateur à être plus agressif sur l'interrogatoire afin d'obtenir ce qu'il veut.

Les sous-classes de **Meta_Variable** sont les suivantes :

(i) **Author_Point_of_view:**

Qui perçoit ou mesure la variable ? Les variables mesurées dépendent parfois beaucoup du regard de celui qui la mesure comme les ressentis ou la qualité de vie. Certaines échelles « objectives » peuvent avoir une concordance moyenne voire médiocre comme l'OMS ou l'ASA. Lorsqu'il existe plusieurs points de vue, il peut exister un biais de mesure qui sera différentiel ou non. Par exemple, il y a une différence parfois importante entre ce que rapporte un enfant et ses parents.

(ii) **Decision-makers:**

Qui décide ou agit ? Celui qui prend la décision est une variable pour laquelle nous n'attendons pas forcément le nom du décideur mais par exemple les médecins généralistes versus les médecins spécialistes, les juniors versus les séniors.

(iii) **Method of Measurement**:

Comment la variable a-t-elle été mesurée ? La méthode de mesure ne veut pas forcément dire examen clinique versus questionnaire, mais aussi examen clinique complet versus incomplet.

(iv) **Location:**

Où prend place la variable ou l'action? La localisation correspond aux variables relatives aux lieux comme l'hôpital versus ville, ou le code postal. Dans les guides de reporting la localisation est celle de l'étude, par exemple étude multicentrique dans le centre A, B, C, D, E. Par exemple, 'Durée de l'hospitalisation' *hasLocation* 'Surgery' hasPossiblevalue 'constant'. Dans certaines circonstances la variable 'location' peut biaiser la relation entre exposition et outcome comme dans le Berkson's bias (collider bias) (Westreich D et al., 2012). Cette variable location ne correspond plus à 'où se déroule l'action' mais à une variable location vers laquelle on va, à cause d'une maladie. Par exemple, lorsqu'on a un diabète déséquilibré, on a plus de risque d'être hospitalisé. L'hospitalisation est ici un traitement (soin hospitalier vs soin de ville), et cette hospitalisation se déroule à l'hôpital (Méta_Variable Location).

(v) **Period:**

Quand l'action prend-elle place ? Dans les guides de reporting, les périodes d'intérêt sont la période : de recrutement, de suivi, d'exposition ou non exposition, avant la découverte d'une maladie, minimal d'exposition. Pour une variable donnée, cela permet de spécifier si par exemple une chirurgie se déroule pendant une garde la nuit versus en journée ou si elle n'est pratiquée que depuis 2015. Pendant la pandémie de coronavirus, les critères d'éligibilité pour la réanimation, les traitements recommandés et la mortalité ont variés au cours du temps,

donc selon la période. Le temps minimal d'exposition est une information constante (non variable dans une étude donnée) qui sera exprimée au travers d'un data object property.

(vi) **Collection_method** :

Lorsque la donnée a été produite une première fois, il faut la collecter. La méthode de collection doit être détaillée car elle peut varier au cours du recueil ou selon les protagonistes qui participent à l'extraction. Si les règles suivies ne sont pas les mêmes, les différences observées entre deux groupes peuvent simplement être dues à cette méthode. L'utilisation des bases de données issue du recueil PMSI (Programme de Médicalisation des Systèmes d'Information) est un bon exemple : les techniciens d'information médicale lisent les comptes rendus et le dossier informatisé du patient afin de définir des codes de diagnostics. Selon les maladies, l'évolution des règles au cours du temps, les techniciens et le centre la concordance est parfois mauvaise. Cette classe est à différencier de **Author_Point_of_View** qui correspond à la première fois où la donnée est créée. On peut dire qu'avant d'être utilisée, la donnée doit être créée puis collectée, elle traverse donc deux filtres ou deux interprétations avant l'analyse.

(vii) Missing value reasons fait partie des Méta_variable.

Exemple récapitulatif :

La variable « taille de la nécrose cérébrale » est mesurée avec une IRM (**Method_of_Measurement**) prescrite par le neurologue (**Decision-maker**), au CHU (Location), durant la journée (**Period**), puis analysé par un radiologue (**Author_Point_of_View**) et les données du compte rendu sont extraites par un algorithme qui repère la taille avec de la fouille textuelle automatique (**Collection_Method**). Cette variable *hasMissing_value_reasons* taille de la nécrose cérébrale NA (i.e., il existe une raison derrière la production de valeurs manquantes de cette variable) (figure 45).

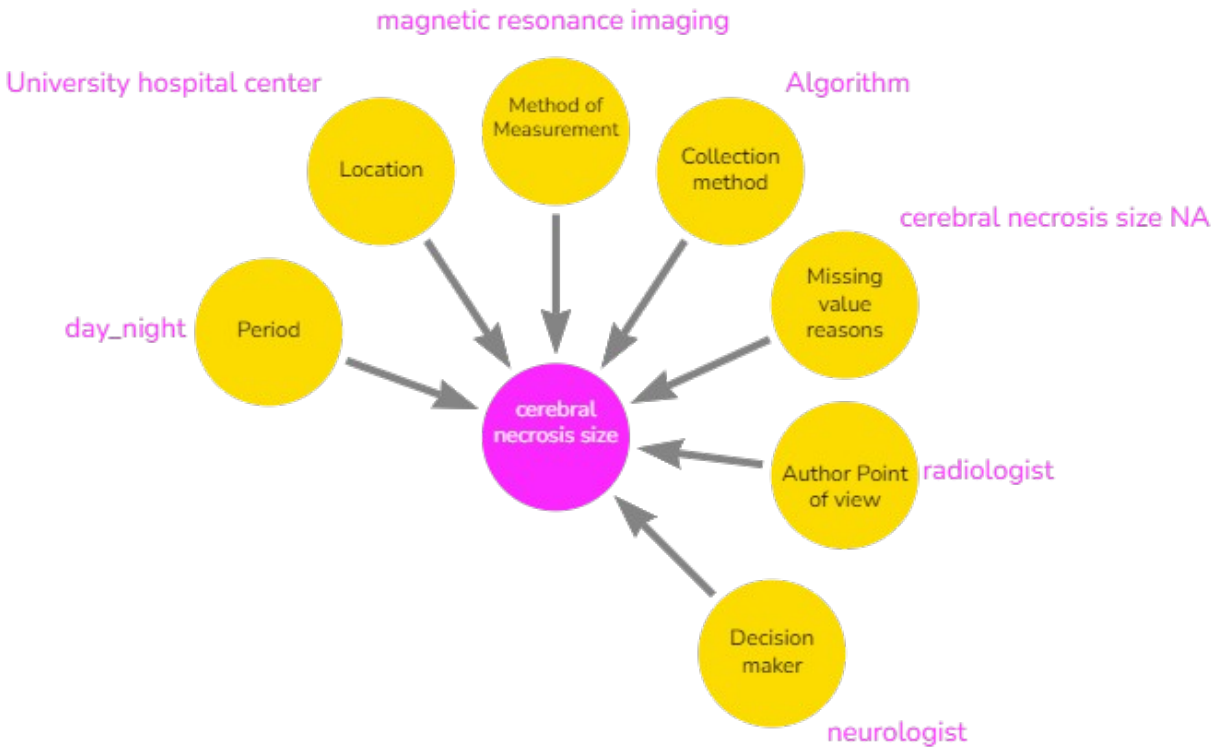


Figure 45. Exemple méta_variable.

Les variables implicites et les variables théoriques : causes nécessaires et conséquences systématiques

Les causes sont nécessaires lorsqu'elles doivent toujours précéder un effet. Les conséquences sont systématiques lorsqu'elles succèdent toujours un évènement. Le but est de rendre explicite ces variables afin de les prendre en compte, au besoin, lors de la construction et des inférences.

Dans l'« ontology of causation », (Galton A, 2012) représentant : (i) états, (ii) évènements et (iii) processus nécessaire ou autorisant la réalisation d'une action donnée ; une porte doit être déverrouillée pour pouvoir être ouverte en tournant la poignée. En partant de ce postulat, j'ai créé une série de classes de variables théoriques regroupant les grandes catégories d'exposition et de critère de jugement. A partir de celles-ci, j'ai formalisé sous la forme de classes les variables nécessaires à leur réalisation et des instances (covariables) génériques de ces classes nécessaires ont été créées. Puis j'ai créé un ensemble de règles basées sur les classes théoriques pour mobiliser des variables nécessaires adaptées à la situation. Ainsi, pour une variable théorique donnée un ensemble de variables nécessaires à sa réalisation seront reliées entre elles et à la variables théoriques grâce aux inférences. Par

exemple si une instance d'Outcome est classée comme une condition, l'inférence de la règle SWRL suivante permet de construire automatiquement un diagramme (Figure 46) :

Outcome(?x) ^ Condition(?x) -> isCauseof(exist outcome, available outcome) ^ isCauseof(available outcome, investigated outcome) ^ isCauseof(available decision maker outcome, decision maker outcome) ^ isCauseof(decision maker outcome, investigated outcome) ^ isCauseof(investigated outcome, measured outcome) ^ isCauseof(exist method of measurement outcome, available method of measurement outcome) ^ isCauseof(available method of measurement outcome, indicated method of measurement outcome) ^ isCauseof(indicated method of measurement outcome, prescribed method of measurement outcome) ^ isCauseof(prescribed method of measurement outcome, method of measurement outcome) ^ isCauseof(method of measurement outcome, measured outcome) ^ isCauseof(measured outcome, ?x)

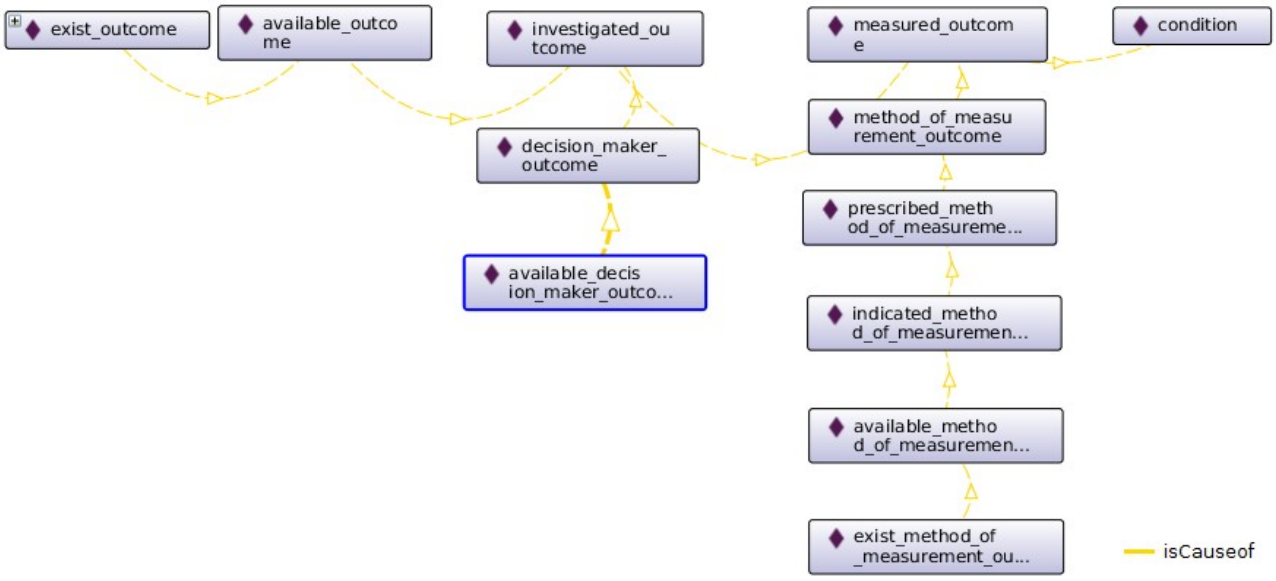


Figure 46: Pre-included instances connected with 'isCauseof' object property when Pellet reasoner is activated (OntoGraf)

Définition des variables théoriques

Ces classes permettent de classer n'importe qu'elle nouvelle variable comme une classe théorique. Cela permet de créer des règles génériques applicables à un sous type de classe théorique. Ainsi, elles permettent de mobiliser les variables nécessaires adéquates à la situation. Initialement, elles concernent uniquement les instances de la classe exposition et celles de la classe critères de jugement (outcome), non les covariables (covariate). Cependant, elles peuvent tout de même être utilisées pour classer les covariables. Les sous-classes de **Theoretical_Variable** sont les suivantes (figure 47) :

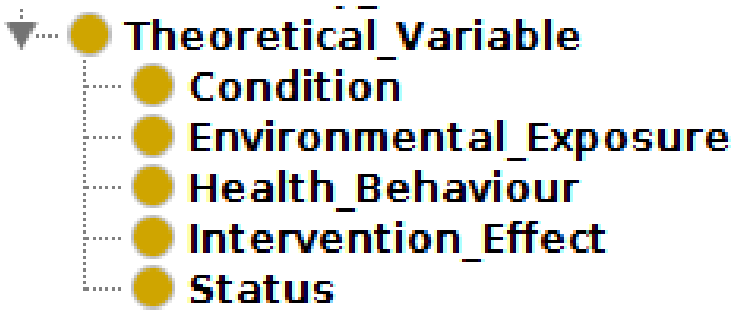


Figure 47: Hiérarchie des classes de variables théoriques

(i) **Condition** (National Cancer Institute Thesaurus: **Disease, Disorder or Finding** de Coronado S et al., 2009) : Une variable sera classée comme **Condition** lorsque la question de recherche porte sur l'effet de la maladie diagnostiquée. Ainsi, si le critère de jugement est le cancer du poumon et que la question porte sur l'impact du tabac, alors le cancer sera classé comme une condition. Si l'exposition est l'effet de la maladie diabète diagnostiquée sur la survenue d'une infection alors le diabète est classé **Condition**.

(ii) **Environmental_Exposure** (from Human Health Exposure Analysis Resource (HHEAR) ontology (Viet SM et al., 2021)) :

(iii) **Health_Behavior** (regroupe 4 classes de HHEAR : **Diet and Nutrition**; **Alcohol, Tobacco and Illicit Drug Use**; **Physical Activity and Fitness**; **Sleep Characteristic**) :

(iv) **Intervention_Effect** (HHEAR classes : **Prescription Medication and Dietary Supplements**, Medical Action Ontology: **medical action (**https://github.com/monarch-initiative/MaxO)) : Une variable sera classée comme **Intervention_Effect** si on s'intéresse à l'effet de la molécule effectivement prise par le patient ou d'une procédure chirurgicale réalisée.

(v) **Status** (regroupes 5 classes de HHEAR **Demographic**, **Anthropometry**, **Dead**, **Socioeconomic Status**, **Medical History**) : Une variable sera classée comme un **Status** si on s'intéresse juste au fait d'être considéré comme malade ou sportif. En effet, le coronavirus n'a pas formellement d'impact au niveau social, cependant le statut covid positif entraîne une éviction voire un confinement de la personne. Lorsque la question de recherche est spécifiée, il faut clairement différencier l'effet de la maladie diagnostiquée de l'effet du diagnostic de la maladie. Cette classe théorique fait partie de la famille des classes dites implicites au même titre que les causes nécessaires développées plus bas. Alors que les causes nécessaires implicites précèdent la variable d'intérêt, le statut est la conséquence implicite de celle-ci.

Les variables théoriques n'ont pour l'instant pas de sous classes. Ces classes sont retrouvées dans deux ontologies et une terminologie que sont HHEAR, Medical Action

Ontology et NCIT. L'ontologie HHEAR rassemble tout l'exposome et permettrait d'aider des implémentations d'études sur les expositions à partir d'études existantes.

Définition des variables nécessaire

Quel que soit la classe théorique de l'exposition ou du critère de jugement, l'instance doit exister et être disponible (visible). Une fois qu'elle est visible (i.e., diagnostiquée ou mesurée ou prescrite), le patient possède le statut correspondant (e.g., statut traité ou statut malade).

(i) Pour la réalisation de **Intervention_Effect** l'intervention doit exister, être disponible, indiquée, prescrite par un médecin disponible, délivrée, le patient doit adhérer au traitement ce qui entraîne l'effet de l'intervention mais aussi le statut « sous intervention ».

(ii) Pour la réalisation de **Condition**, elle doit exister, être disponible dans le sens « diagnostiquable », investiguée par un médecin lui-même disponible, mesurée avec une méthode de mesure qui existe, est disponible, prescrite pour que la maladie soit diagnostiquée.

(iii) Pour la réalisation d'une exposition environnementale, elle doit exister et être disponible dans le sens peut entraîner une exposition. Si l'uranium reste dans sa grotte il n'est pas disponible pour entraîner une exposition. Pour la faire exister, il faudra une méthode de mesure. En l'absence d'outil de mesure existant ou de démarche de recherche, il n'est pas possible de rendre visible cette exposition.

(iv) Pour la réalisation d'un comportement de santé, il faut les mêmes éléments que pour une intervention. C'est une intervention auto prescrite, délivrée (e.g., un bureau de tabac, une salle de sport) et à laquelle le patient adhère.

(v) Pour le statut, les variables nécessaires sont celles des autres classes théoriques. Le sexe biologique, l'âge, le niveau socio-économique ne nécessitent pas de variable nécessaire. C'est un état de faits.

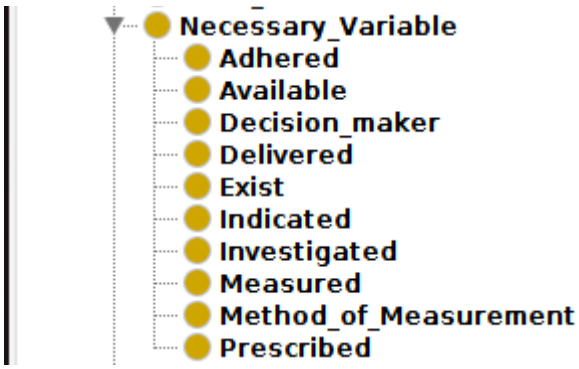


Figure 48: Hiérarchie des classes nécessaires

La classe **Necessary_Variable** regroupe toutes les variables nécessaires pour la réalisation d'une instance **Exposure_stressor** ou **Outcome.** Les huit sous-classes de Necessary_Variable sont les suivantes (figure 48) :

- **Exist** est le prérequis le plus évident nécessaire à la réalisation de la chose.

- **Available** correspond au fait que pour se réaliser, la chose doit se rendre disponible, soit accessible pour une intervention chirurgicale, soit présent en pharmacie pour un médicament, pour une maladie il faut qu'elle puisse être 'diagnostiquable'. En résumé, il faut que la réalisation de la chose soit 'logistiquement' possible en dehors la volonté du patient.

- **Indicated** : les réalisations d'interventions ne sont pas aléatoires et doivent être indiquées soit pas des symptômes du patient soit par des caractéristiques (e.g., l'âge et le sexe pour le dépistage du cancer du sein).

- **Prescribed** : beaucoup d'interventions dépendent d'une prescription d'un professionnel de santé.

- **Delivered** : en ce qui concerne l'intervention, il est nécessaire qu'elle soit délivrée pour se réaliser.

- **Investigated** : pour être diagnostiquée, il doit nécessairement y avoir une recherche qu'elle soit orientée ou non vers la maladie qui sera diagnostiquée. Elle est amorcée par un professionnel qui va chercher à mesurer une chose.

- **Adhered** : le patient s'il peut choisir le fait de subir une intervention ou pas doit adhérer.

- **Measured** : la chose doit être mesurer avec une méthode de mesure (Method_of_Measurement) elle-même devant exister, être disponible, indiquée, délivrée et acceptée par le patient. Attention, toutes les variables doivent être mesurées. Ici, on s'intéresse au fait que pour devenir visible il faut qu'il y ait une démarche de mesure. C'est à dire qu'il existe beaucoup d'éléments invisibles qui nécessitent d'être mesurés pour les détecter. Quelqu'un qui fume c'est visible, la pollution dans l'air non.

Ces instances sont pour la plupart des variables non mesurées (datatype property isMeasured FALSE). Cependant, elles permettent d'inférer si une covariable est une variable de confusion ou non. L'object property isCauseof est utilisé pour relier ces instances alors qu'une relation causale plus formelle du type 'permet' ou 'est nécessaire pour' aurait été plus approprié sémantiquement parlant.

Apports des causes nécessaires et métavariables

Voici quatre exemples théoriques dans lesquelles les variables nécessaires ont un apport, decision-maker et method of measurement sont aussi des causes nécessaires :

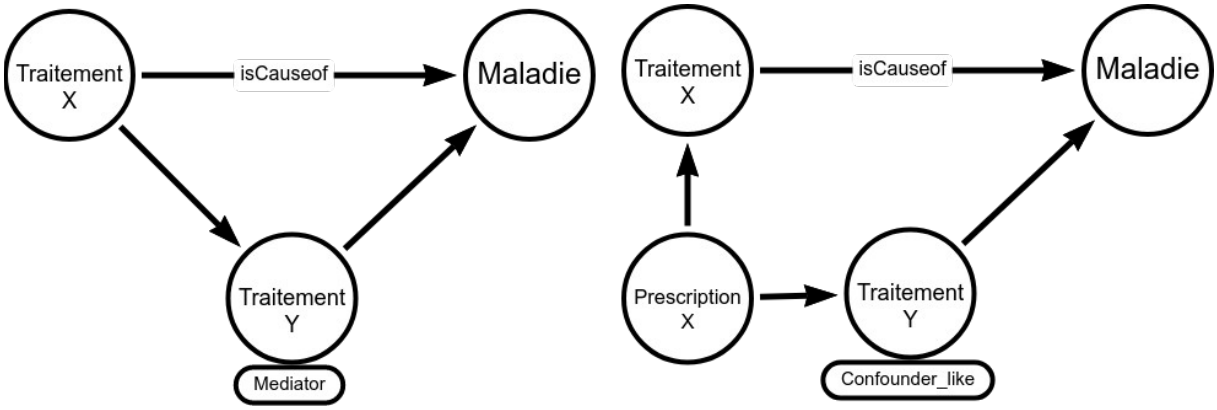


Figure 49: Mediator to Confounder-like (prescription).

Voici ici un premier exemple théorique de l'impact des variables nécessaires sur la sélection des variables. Soit un jeu d'instances suivant : Traitement X (l'exposition), Traitement y (covariable) et une Maladie. Nous savons que :

- la prescription du 'Traitement X' mène généralement à la co-prescription du 'Traitement Y' ('Traitement X' *isCauseof* 'Traitement Y')
- 'Traitement Y' cause la maladie ('Traitement Y' *isCauseof* 'Maladie')

Nous voulons savoir si le traitement X cause la maladie. La question est donc : est ce que le traitement Y doit être sélectionné dans le modèle ou non. Dans la figure 49, le premier diagramme construit sans la conscience des variables nécessaires arrive à la conclusion que le traitement Y est une variable de médiation et donc qu'il ne doit pas être sélectionné. Dans le deuxième diagramme le nœud prescription x est automatiquement créé et permet de se rendre compte que le traitement Y est un **Confounder_like** et doit donc être sélectionné.

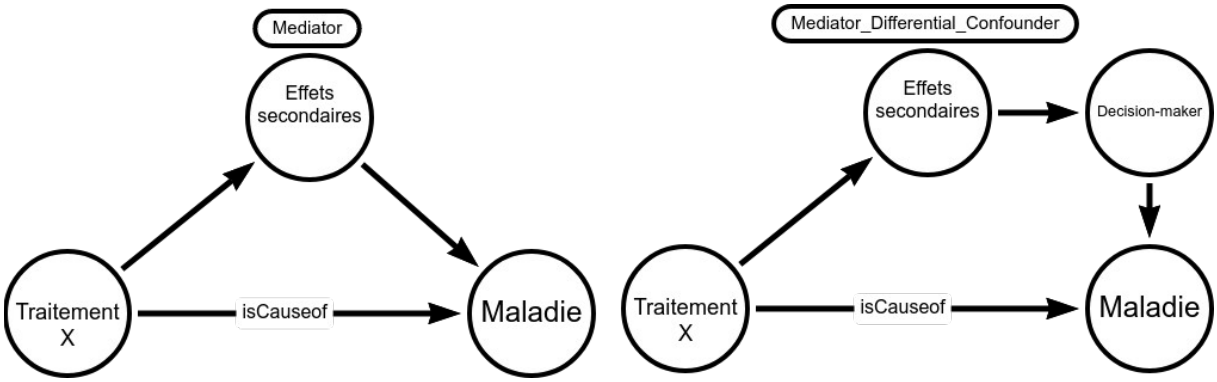


Figure 50. Mediator to Mediator Differential Confounder.

Deuxième exemple théorique de l'impact des variables nécessaires sur la sélection des variables : soit un jeu d'instances suivant : Traitement X (l'exposition), Effets secondaires (covariable) et une Maladie. Nous savons que :

- le ‘Traitement X’ a des effets secondaires ‘Effets secondaires’ (‘Traitement X’ *isCauseof* ‘Effets secondaires’)
- ‘Effets secondaires’ cause des rencontres plus fréquente avec le soin, ce qui permet le diagnostic précoce de la Maladie (‘Effets secondaires’ *isCauseof* ‘Maladie’)

Nous voulons savoir si le traitement X cause la maladie. La question est donc : est ce que la variable ‘Effets secondaires’ doit être sélectionnée dans le modèle ou non. Dans la figure 50, le premier diagramme construit sans la conscience des variables nécessaires arrive à la conclusion que Effet secondaire est une variable de médiation et donc qu’elle ne doit pas être sélectionnée. Dans le deuxième diagramme le nœud decision-maker est automatiquement créé et l’effet secondaire est toujours une variable de médiation. Cependant d’un point de vue logique on comprend bien que l’effet secondaire n’est en rien responsable de la survenue de la maladie, simplement, le diagnostic est plus précoce. L’impact est particulièrement important dans le cadre des maladies qui apparaissent lentement ou qui deviennent ‘bruyantes’ sur le tard comme certain cancers ou démences. L’effet secondaire doit donc être pris en compte pour ne pas raccourcir artificiellement la découverte de la maladie. Une nouvelle classe de facteur de confusion appelée **Mediator_Differential_Confounder** est créée. La règles SWRL suivante permet l’inférence :

Inverse_Directed_Relation(?n, ?m) ^ Necessary_Variable(?n) ^ Mediator(?m) -> Mediator_Differential_Confounder(?m)

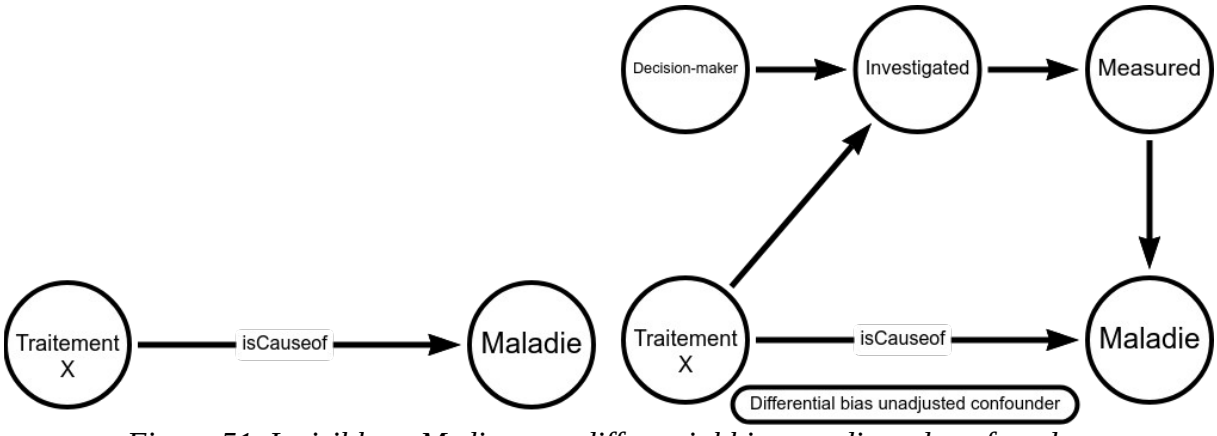


Figure 51. Invisible to Mediator to differential bias unadjusted confounder.

Troisième exemple théorique de l’impact des variables nécessaires sur la sélection des variables : soit un jeu d’instances suivant : Traitement X (l’exposition) et une Maladie. Nous savons que :

- sous ‘Traitement X’ il est plus facile de diagnostiquer une Maladie.

Nous voulons savoir si le traitement X cause la maladie. La question est donc : est ce qu'il peut exister un biais qui confondrait l'effet causal entre traitement X et la maladie. Dans la figure 51, le premier diagramme construit sans la conscience des variables nécessaires arrive à la conclusion qu'il n'y a rien qui pourrait confondre l'effet causal entre Traitement X et la Maladie. Dans le deuxième diagramme plusieurs nœuds sont créés. Le traitement X cause la Maladie par l'intermédiaire d'une variable nécessaire. Malheureusement, cette variable n'est pas mesurée et il n'existe aucun proxy de celle-ci. L'exposition est par essence vectrice de plus de Maladie sans qu'elle en soit la cause directe. Une nouvelle classe est créée afin de rappeler ce biais de mesure inhérent au fait d'être traité par X : **Differential bias unadjusted confounder**. Les règles SWRL sont les suivantes :

OntoBioStat:Necessary_Variable(?c) ^ OntoBioStat:Exposure_stressor(?e) ^ OntoBioStat:isCauseof(?e, ?c) -> OntoBioStat:Differential_Bias_Unadjusted_Confounder(?e)

OntoBioStat:Method_of_Measurement(?c) ^ OntoBioStat:Exposure_stressor(?e) ^ OntoBioStat:isCauseof(?e, ?c) -> OntoBioStat:Differential_Bias_Unadjusted_Confounder(?e)

Une dernière situation pourrait impliquer uniquement les variables nécessaires de l'exposition et du critère de jugement. Là encore, le biais ne pourrait pas être corrigé.

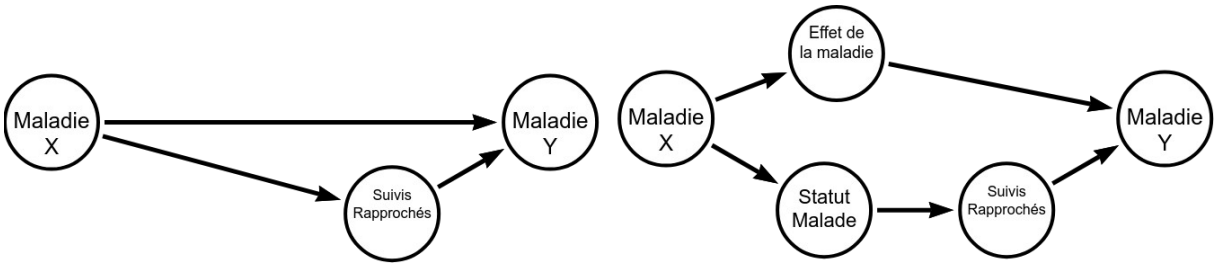


Figure 52. Mediator to Confounder-like (statut).

Quatrième exemple théorique de l'impact des variables nécessaires sur la sélection des variables : dans le chapitre d'ouverture sur la recherche biomédicale, il est fait mention du biais de sélection. Ce biais de sélection inclut les biais de suivi, c'est-à-dire les biais survenant à cause d'un suivi différent dépendant du statut exposé ou non exposé d'un patient. Soit un jeu d'instances suivant : Une maladie X (l'exposition), une maladie Y (l'outcome) et le suivi rapproché avec éducation thérapeutique, surveillance et soins réguliers (covariable). Nous savons que :

- Les personnes avec la maladie X ont un suivi rapproché chez leur médecin
- Les personnes avec la maladie X auraient plus de maladie Y
- Les personnes avec un suivi rapproché chez leur médecin ont moins de maladie Y

Grâce à la nuance entre l'effet de la maladie et le statut malade, il est possible de mettre en évidence que le suivi rapproché est bien une variable de confusion. En effet, c'est le statut malade qui entraîne un suivi rapproché et non l'effet de la maladie (Figure 52).

Incertitude des relations causales

L'utilisation de la connaissance peut engendrer des erreurs surtout si le degré de certitude des faits qui sont spécifiés lors de la construction d'un diagramme n'est pas représentée et donc non pris en compte lors de la sélection des variables. Les graphes orientés acycliques reposent sur une connaissance qui est à la fois consolidée, changeante, incertaine, hypothétique et incomplète : (i) Consolidée car il existe des connaissances médicales qui forment un socle immuable comparable à l'assertion « la Terre est ronde », (ii) Changeante car les indications d'un traitement même si elles sont sûres et consolidées à un instant peuvent changer l'année d'après ou dans un autre centre, (iii) Incertaine car une partie de la connaissance reste encore à l'épreuve de la médecine basée sur les preuves, (iv) Hypothétique car une partie de la connaissance sont des suppositions ou spéculations n'ayant pas été confirmées par un article scientifique mais seulement par une poignée de cliniciens du domaine, et (v) Incomplète car nous ignorons ce que nous ne savons pas. Incertaine et Hypothétique peuvent entraîner une variabilité pour une étude donnée alors que Changeante sur deux études semblables à un moment différent ou une même étude lorsqu'elle se déroule sur une période étendue ou dans des lieux très différents. Cependant, ces limites peuvent être contrebalancées par l'utilisation d'analyses de sensibilité, c'est-à-dire en proposant différents graphes pour une même étude.

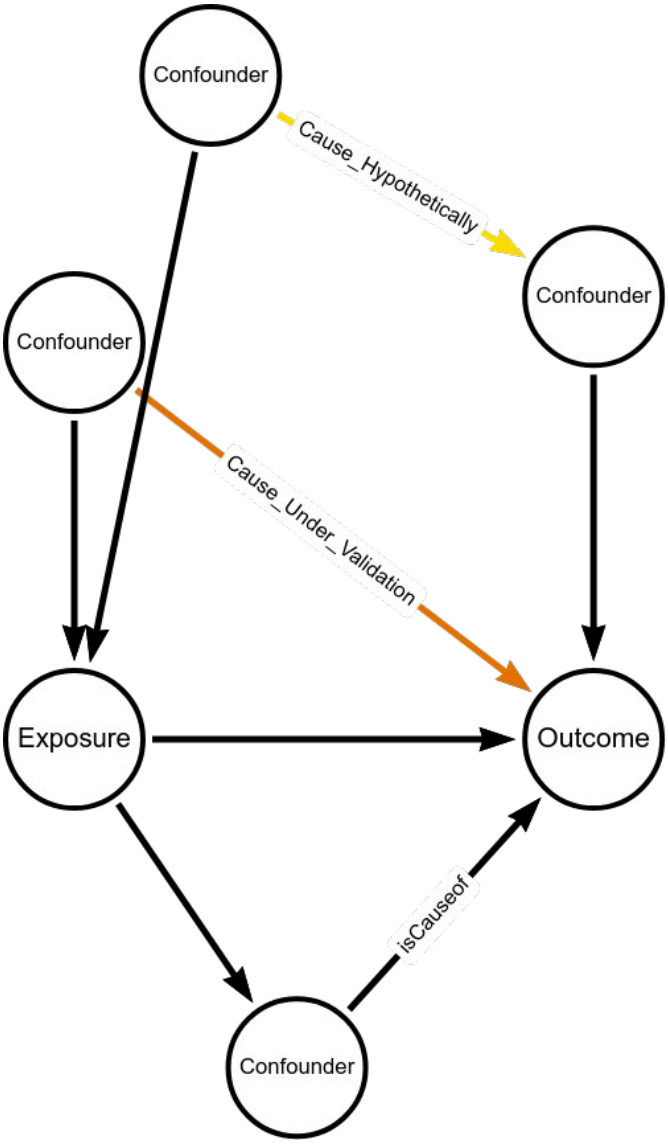


Figure 53: Liens causaux hypothétiques (jaune) et en cours de validation (orange).

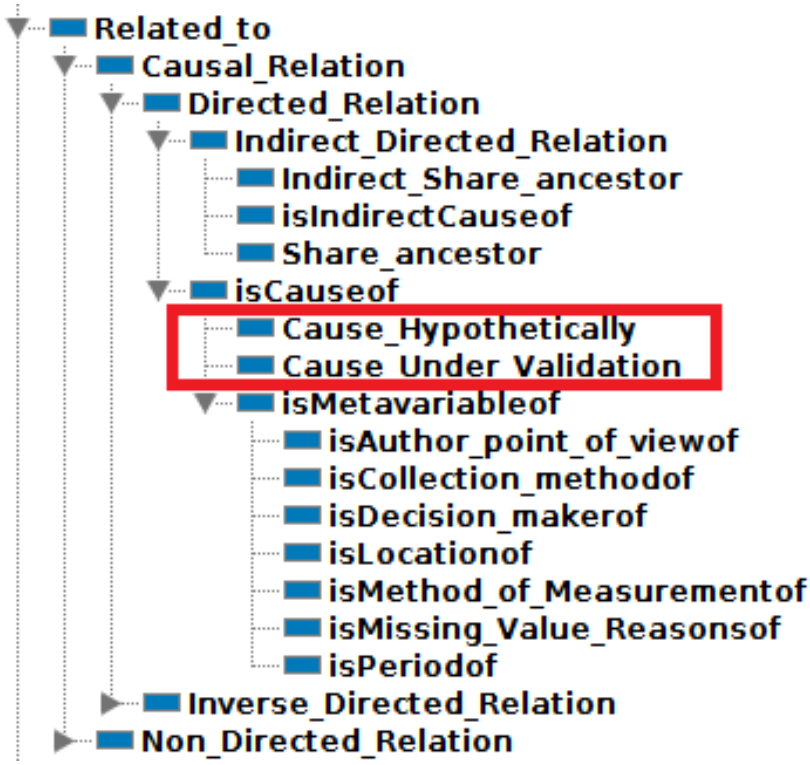


Figure 54: Nouvelles relations sur l'incertitude intégrées à la hiérarchie des relations causales

Dans OntoBioStat, cette incertitude a été représentée par de nouvelles relations causales descendant de la relation isCauseof (figure 54) :

Cause_Hypothetically : cet object property représente les relations causales qui sont pures spéculations.

Cause_Under_Validation : correspond aux relations causales qui ont été retrouvées dans les articles mais qui nécessitent encore une confirmation avant de devenir un savoir immuable (connaissances incertaines).

Les autres types de liens causaux (immuables et changeants) ne seront pas représentés. L'utilisateur pourra de cette manière savoir si une relation hypothétique est impliquée dans l'inférence d'une variable de confusion ou non. Ainsi, il pourra préférer un jeu de variable minimum qui exclut les liens causaux hypothétiques et proposer des analyses de sensibilité le cas échéant. Par exemple, dans la Figure 53, il y a quatre variables de confusion dont trois dues à la présence de lien cause hypothétique ou en cours de validation. Il sera possible de proposer 2 à 3 modèles différents incluant ou non ces variables supposées confondre l'effet causal entre exposition et critère de jugement.

Data properties

Les datatype properties permettent d'ajouter des informations soit sous forme de littérale, soit numérique, soit catégorielle concernant les variables. Les différentes datatype properties ont un point commun : elles ne varient pas pour une variable donnée. Par exemple, la variable 'tabac en paquets années' est une variable numérique. Elle ne peut pas être à la fois numérique et catégorielle. Si la variable 'tabac en paquets années' est recodée en fumeur « oui » versus « non » alors cela devient une autre variable catégorielle qui pourrait avoir pour nom 'fumeur actuel'.

Data properties générales

(i) has**Definition.** Une même variable dans deux études différentes peut avoir une définition différente donc une définition devrait être spécifié pour chacune. Dans les guides de reporting il faut fournir les critères diagnostics qui sont intrinsèquement liés à la définition de la variable (i.e.,, diagnostic criteria).

(ii) has**Type** (dichotomous, ordinal, polychotomous, quantitative). Statistics Ontology (http://statoontology.org) and Human Physiology Simulation Ontology décrivent ces entités comme des sous classes de **Variable** (Gündel M et al., 2013).

(iii) has**Possible_value**. Elle représente les valeurs possibles pour une variable donnée. ("Constant", "Bounded", "Complete","Incomplete"). "Constant": Le fait qu'une variable soit en réalité une constante (Path_Modifier) est souvent omis car ce ne sont pas des variables à proprement parler. Cependant, une variable constante est un Path_Modifier. Donc, si cette constante est un collider ou un mediator cela va biaiser le vrai effet causal. "Bounded" (Panov P et al., 2014; Panov P et al., 2016) est encore plus subtil que constante, car la variable va varier mais pas dans son domaine complet de valeurs. Ces covariables peuvent être identifiées via les critères d'éligibilité d'une étude donnée (e.g., age >18). "Incomplete": Incomplète signifie une variable qui a déjà perdu de l'information car elle a déjà subi un processus de transformation (e.g., âge de 0 à 75 transformé en variable catégorielle: [0-10][11-20][21-40] …). Des transformations inappropriées peuvent amener à des pertes qui sont préjudiciables lorsque l'on veut corriger les biais en sélectionner les variables.

"Complete": C'est une variable qui n'est ni incomplète, constante ou bounded.

Dans les guides de reporting, il faut définir le découpage éventuel d'une variable quantitative (i.e., category boundaries for continuous variable). Les deux dernières datatype properties permettent d'avoir cette information.

(iv) **Data_source.** Elle correspond aux types de jeu de données (e.g., "registry", "electronic health report", ...). Data source est indispensable pour apprécier la qualité des données, leur

complétude et l'extrapolation possible. On préfèrera une variable qui a une source de données connue comme fiable pour la sélection des variables. Data source fait partie des items requis par les guides de reporting.

Data properties concernant les données manquantes

(i) **mayhaveMissing_value.** Les covariables collectées peuvent avoir des données manquantes (e.g., il existe une imputation simple par 0 par défaut dans les bases de données médico administratives).

(ii) has**Missing_value_amount.** Elle correspond au pourcentage de données manquantes pour une variable donnée. Dans le cas d'une analyse en cas complet (*complete case analysis* ou *leastwise*), inclure des covariables avec données manquantes dans un modèle multivarié entraîne la perte d'observations quand il y a au moins une variable avec une valeur manquante. Une covariable avec un plus bas pourcentage de données manquantes sera préférée.

Exemple récapitulatif :

La variable taille de la nécrose cérébrale a pour définition *literal*, est une variable de type quantitative, has Possible_value complete, has data source electronic health report, mayhaveMissing_value TRUE and hasMissing_value_amount 15 % (Figure 55).

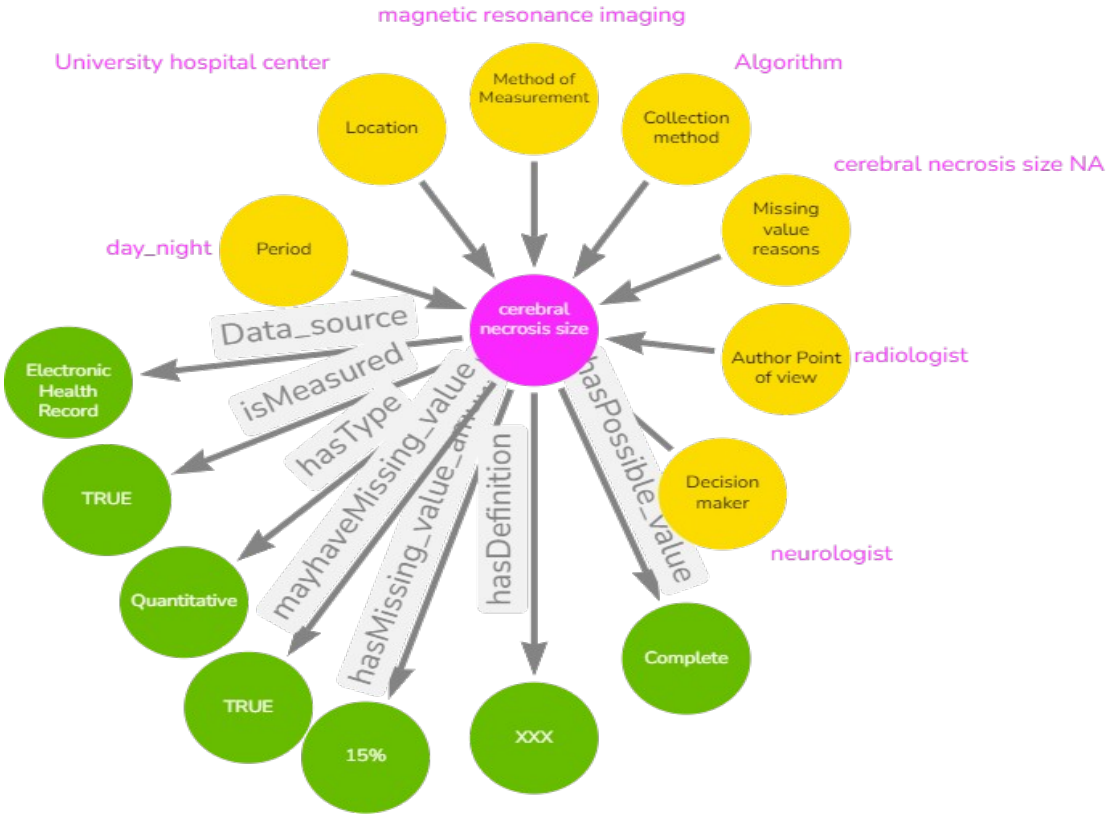


Figure 55. Exemple d'utilisation des dataobject properties.

Data properties concernant la chronologie

La représentation de la chronologie est très importante que ce soit dans la causalité ou dans la prédiction. Cependant, dans la causalité les object properties *isCauseof* impliquent *isBefore* ce qui est suffisant. De plus, contrairement à la prédiction, dans la causalité la sélection des covariables n'est pas restreinte par la ligne du temps. C'est ainsi qu'il est possible de sélectionner des variables qui se passent après l'exposition. Allen algebra ou Time Event Ontology n'ont pas été considérées dans OntoBioStat (Li F et al., 2020).

Il existe un biais chronologique qu'il est possible de mettre en évidence avec quelques informations sur les variables. Quatre datatype properties en rapport avec le temps furent créées pour contrer le biais de temps immortel (Suissa S et al., 2008) : (i) **hasMinimum_Follow_up_Before_Outcome ;**(ii)**hasMaximum_Follow_up Before Exposure,** (iii) **hasInduction ;** et (iv) **hasLag.** Le biais de temps immortel correspond à la durée pendant laquelle les exposés ne peuvent pas présenter de critère de jugement (e.g., participants cannot die). Même si ce biais peut être représenté par un diagramme causal (Mansournia MA et al., 2021), le biais n'est pas contrôlé par les ajustements. La durée maximale entre t_0 (début du suivi ou du traitement) et le statut exposé ne doit pas excéder la durée minimale entre t_0 et la survenue du critère de jugement. En outre, si le critère de jugement est la survenue d'une maladie, il risque d'y avoir un délai entre le moment où les

premiers symptômes apparaissent et son diagnostic. De même, il peut exister un temps d'exposition minimal avant qu'un individu soit considéré comme exposé. Le temps d'induction et le lag avant diagnostic sont deux éléments demandés par les guides de reporting.

2.2.5. Réajustements de l'ontologie

La création de la classe **Inferred variable** sous classe de **Variable** qui contient les classes **Explicative_Variable** et **Outcome** a été nécessaire. Inferred a été créée pour regrouper les inférences concernant les variables dans l'ontologie et séparer les métavariables, les variables théoriques et nécessaires. Annulation de la disjonction entre **Variable** et **Path_Modifier** et déplacement de la disjonction entre **Path_Modifier** et **Inferred_Variable**. En effet, une **Métavariable** peut être constante et donc être un **Path_Modifier**.

2.2.6. Résumé de OntoBioStat

OntoBioStat, à son stade actuel contient 57 classes, 35 object properties, 11 data properties, 30 instances génériques et 33 règles SWRL (Tableau 5).

Tableau 5: Métriques de OntoBioStat telles que rapportées dans l'outil Protégé

Axioms count	453	Annotation assertion	82
Declaration axioms	134	Logical axioms	237
Class count Object Properties count Individuals count Data property count Annotation property	57 35 30 11 1	Equivalent classes axioms SubClassOf axioms Disjoint classes axioms SubObjectPropertyOf InverseObjectProperties DisjointObjectProperties SymmetricObjectProperty AsymmetricObjectProperty IrreflexiveObjectProperty ObjectPropertyDomain ObjectPropertyRange ClassAssertion DataPropertyDomain DataPropertyRange SWRL rules	8 55 4 27 3 1 7 7 7 2 2 60 10 11 33

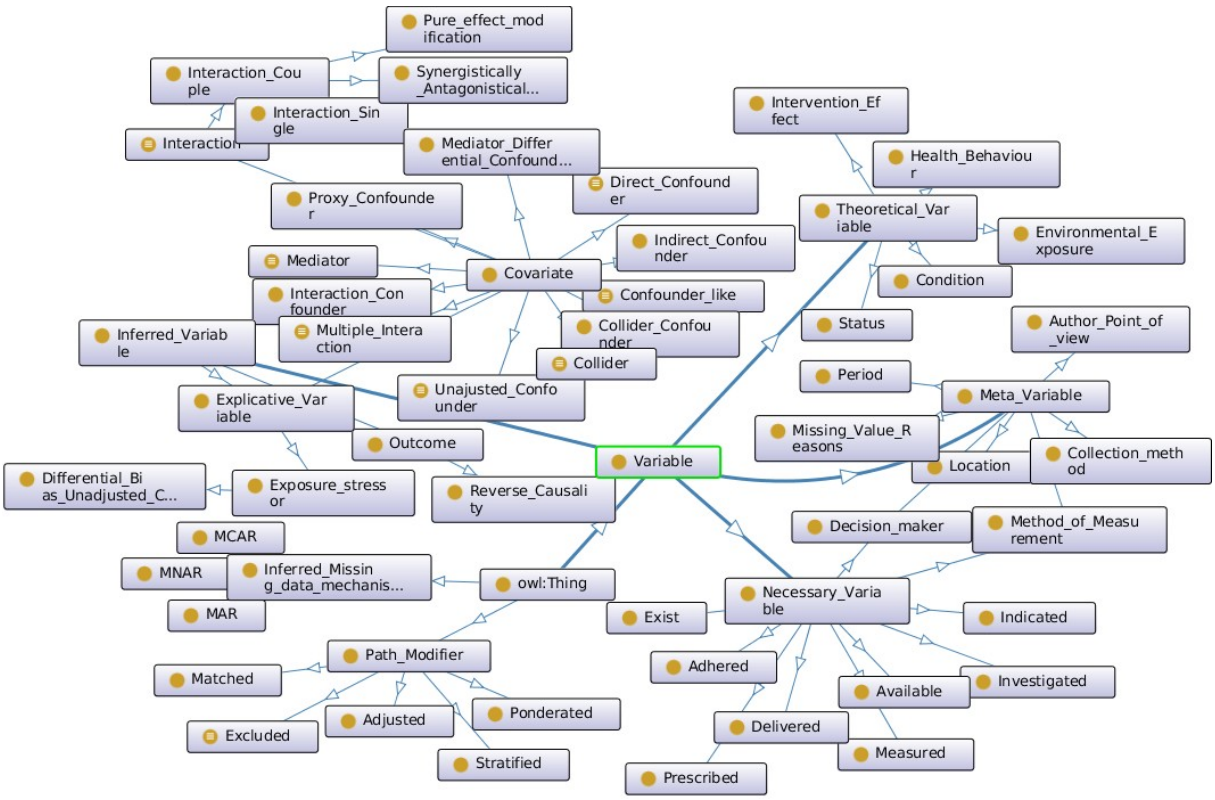


Figure 56: OntoBioStat représentée avec toutes ses classes

Les classes de OntoBioStat peuvent être séparées en entités inférées (via les axiomes et les règles SWRL) et entités spécifiées (le cadre de construction).

Concernant les entités spécifiées, pour les classes : la distinction Exposure, Outcome Covariable est retrouvée ; pour les object properties : isCauseof et ses sous object properties, les relations signées, les interactions, l'absence de cause et les relations causales vers les données manquantes ; et pour les datatype properties sept informations complémentaires.

Dans les classes inférées on retrouve :

(i) les classes nommées Confounder au nombre de huit qui permettent de répondre à cinq questions de compétence (figure 57) ;

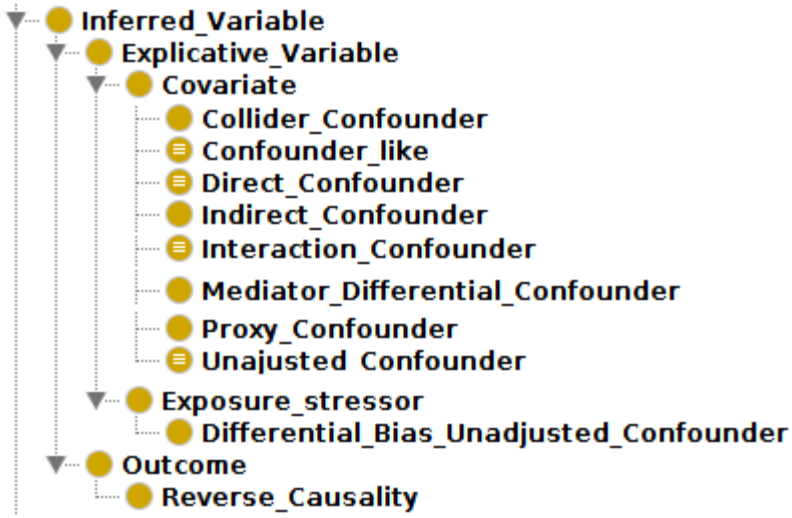


Figure 57: Hiérarchie de classes inférées comprenant les variables de confusion et les biais qu'on ne peut corriger

(ii) les classes **MNAR** et **MAR** qui répondent à une question de compétence ; (iii) la classe **Reverse Causality** qui permet de répondre avec **Diffrential Bias Unadjusted Confounder** à une huitième question de compétence : « Est-ce qu'il existe des biais qu'on ne peut pas corriger ? » (Tableau 6) (figure 57).

Dans les relations inférées on retrouve les object properties qui permettent de détailler la relation entre deux variables et ainsi répondre à la septième question de compétence.

Les huit classes qui correspondent aux variables nécessaires sont inférées mais ne permettent pas de répondre directement aux questions de compétences. Elles servent dans la construction du diagramme et dans les inférences.

Tableau 6: Questions de compétences avec les classes et object properties qui permettent de répondre

1) Faut-il exclure des patients qui ne seraient pas comparables aux autres ?	**Unadjusted_Confounder**
2) Quelles sont les variables qui confondent le vrai effet causal entre exposition et critère de jugement ?	**Confounder_like**, **Direct_Confounder**, **Indirect_Confounder**, **Mediator_Confounder**, **Interaction_Confounder**, **Collider_Confounder**.
3) Existe-t-il une interaction qui pourrait biaiser le vrai effet causal entre exposition et critère de jugement ?	**Interaction_Confounder**
4) Est-ce que le mécanisme responsable de la présence de données manquantes pourrait	**MNAR, MAR**

biaiser l'estimation du vrai effet causal entre exposition et critère de jugement ?	
5) Quelle est la direction des biais causés par des variables de confusion ?	**Decrease, Increase**
6) Existe-t-ilil des variables proxies de variable de confusion ?	**Proxy_Confounder**
7) Quelle type de relation existe entre deux variables ?	Tous les object properties en dessous de related to.
8) Est-ce qu'il existe des biais qu'on ne peut pas corriger ?	**Differential_Bias_Unadjusted_Confounder , Reverse_Causality**

2.3. Validation avec OOPS

L'évaluation avec l'outil en ligne OOPS révèle les erreurs suivantes :

- Aucun des éléments de OntoBioStat n'a de label car le nom des classes est déjà explicite. En effet, c'est une ontologie d'application, le standard pour l'alignement avec d'autres ontologies est donc moins important.

- Les classes Path_Modifier et Inferred_Missing_data_mechanism sont deux classes non connectées aux autres éléments de l'ontologie. Cependant, ils sont utilisés dans les inférences SWRL.

- Des relations inverses d'object properties asymétriques ne sont pas déclarées car il n'y en a pas besoin pour les inférences et l'utilisateur.

- Il n'existe pas de convention de nommage mise à part celle décrite plus haut.

- Le domaine et le co-domaine n'ont pas été précisés dans certaines object properties parce que tout est possible en terme de domaine et de range.

- Les classes Status et Condition sont détectées comme possiblement équivalentes alors que, dans OntoBioStat, il est bien défini que ce sont deux classes différentes. Une condition correspond à un symptôme ou une maladie alors que le statut correspond au fait d'être malade.

3. Discussion

OntoBioStat est une ontologie d'application pour la sélection des variables dans le cadre d'une question de recherche de type causal posée à partir de données provenant d'une étude observationnelle. Sa construction consistait en la création d'un corpus de termes extraits manuellement à partir d'articles scientifiques et une ontologisation guidée par des questions de compétences, des tutoriels de construction de diagrammes causaux et la volonté d'expliciter l'implicite en conservant un formalisme minimal pour ne pas alourdir l'ontologie. C'est un nouveau modèle qui se positionne aux côtés des graphes acycliques et des diagrammes causaux utilisés en épidémiologie. Il n'existe pas à notre connaissance d'ontologie dont le but est la sélection des variables.

3.1. Comparaison aux DAG et aux autres ontologies

Il serait difficile de comparer OntoBioStat aux Ontologies médicales, épidémiologique, ou statistique si le versant causal ou sélection des variables ne sont pas représentés. OntoBioStat sera donc comparée aux ontologies (biomédicales) causales et aux diagrammes causaux que ce soit au niveau de la richesse de la représentation que des inférences possibles. De plus, la place d'OntoBioStat dans le paysage de l'épidémiologie sera discutée.

3.1.1. Causalité(s)

La causalité dans OntoBioStat est représentée grâce à l'object property central isCauseof. Il permet d'inférer treize objects properties, et regroupe divers sub-properties qui connectent les métavariables aux variables et permet de spécifier l'incertitude. Cette représentation respecte la logique qui permet d'inférer la corrélation (Related_to) entre deux variables à partir de la structure causale, et les relations issues des métavariables (spatio-temporelle, les méthodes de mesures ou de collection et le point de vue). En effet, deux variables sont corrélées si A cause B ou B cause A, A et B ont une cause commune, ou A et B ont une conséquence commune constante (cause pouvant être interchangée avec *a pour lieu* ou *a pour période*). Selon le Medical Subject Heading thesaurus '[…] Causes are termed necessary when they must always precede an effect and sufficient when they initiate or produce an effect. […]'. Dans OntoBioStat les causes suffisantes sont identifiées par les object properties *Indication* et les *Contre indication absolues*. Les causes nécessaires implicites sont représentées par des classes de variables nécessaires et automatiquement proposées en fonction des classes théoriques spécifiées.

Les inférences d'OntoBioStat basées sur cette représentation causale, sont variées et permettent de répondre à huit questions de compétences essentielles pour guider la sélection des variables. Ces inférences sont le résultat de règles SWRL ou d'axiomes dont le raisonnement est transparent. De plus, il est possible d'avoir des informations supplémentaires via les datatypes properties qui permettent d'apprécier la qualité des données, leur source et la définition des variables.

OntoBioStat et les diagrammes causaux

OntoBioStat intègre les différentes représentations et types de relations causales retrouvées dans les DAG et diagrammes causaux (i.e., flèches unidirectionnelles, bidirectionnelles, non dirigées). OntoBioStat se distingue de ceux-ci par un cadre de construction enrichi avec des relations hasMetavariable, des relations permettant d'apprécier l'incertitude, les causes suffisantes et nécessaires. Mises à part les relations permettant d'apprécier l'incertitude, les diagrammes causaux peuvent très bien représenter avec une flèche dirigée, tout ce qui correspond à hasMetavariable, causes suffisantes et nécessaires. Cela permet d'avoir une représentation facile à prendre en main et compréhensible par tout le monde. Cependant, il existe des inférences spécifiques basées sur les causes suffisantes (question de compétence numéro 1) et nécessaires (variables de confusion inférées grâce aux variables nécessaires) que les diagrammes causaux ne peuvent donc pas résoudre. Les diagrammes causaux n'autorisent pas le dessin d'un cycle (ils sont acycliques), ce qui empêche de se rendre compte d'un effet causal inverse qu'il est possible d'inférer grâce à OntoBioStat. La dernière question de compétence créée au cours du processus de construction de OntoBioStat : 'existe-t-il des biais qu'on ne peut pas corriger ?' ne peut obtenir de réponse de la part d'un diagramme causal classique, car, ni l'effet causal inverse, ni le biais issu de l'effet médié passant directement de l'exposition à une variable nécessaire ne peuvent être inférés.

L'inférence de proxy passe par la connaissance du statut mesuré ou non de la variable et l'absence de liens causaux multiples entre la variable proxy et le critère de jugement ou l'exposition. Il serait possible de l'inférer à partir d'un diagramme causal via une opération sur graphes, mais la représentation NotCauseof de OntoBioStat sécurise le repérage des proxies en explicitant l'absence de liens causaux avec le critère de jugement ou l'exposition. Concernant la représentation des signes des relations causales, OntoBioStat apporte les contre indications/indications absolues, mais détériore la finesse des relations signées de trois types développées dans VanderWeele TJ et al., 2010. De plus, Increase et Decrease n'étant pas subsumés par isCauseof cela oblige à une double spécification de la part de l'utilisateur.

Par ailleurs, il existait plusieurs façon de représenter les interactions dans les DAG (VanderWeele TJ , 2007; Weinberg CR, 2007 ; Nilsson A et al., 2021; Lopez PM et al., 2019). Pour OntoBioStat, il a été fait le choix d'un autre type de représentation assez lourd qui évoluera peut être vers la création de fusion d'instances (e.g., l'interaction entre A et B devient AB) appartenant à la classe **Interaction**, appartenance spécifiée par l'utilisateur. Le raisonnement sur les données manquantes est souvent réalisé de manière indépendante.

La représentation du mécanisme de génération des données manquantes est possible sur un diagramme causal (Mohan K et al., 2021). Considérant que les biais entraînés par les variables avec données manquantes sont à mettre en évidence pour faire une bonne sélection, OntoBioStat mixe causalité pour la sélection des variables à proprement parler mais aussi causalité pour comprendre les variables les plus biaisées à cause des données manquantes sur une même représentation.

Cette volonté de vision globale de la qualité des données rejoint les data properties qui permettent d'obtenir d'autres informations hors champs de la causalité telle que la source des données, la complétude des variables et leur pourcentage de données manquantes.

Pour finir, il n'existe aucune explication sur les variables à sélectionner mises en exergue grâce à un diagramme causal. Les variables à sélectionner mises en évidence sur un diagramme classique n'explique pas pourquoi elles sont à sélectionnées.

OntoBioStat et les autres ontologies

La hiérarchie des relations de OBOREL, l'ontologie des relations, entre en conflit avec celle de OntoBioStat. Celle d'OntoBioStat pose deux questions : qu'est ce qui peut être inféré avec la relation de base isCauseof ? et qu'est ce qui correspond à un enfant de isCauseof ?. OBOREL, elle, regroupe et représente de manière formelle tout type de relations causales ou non. Cependant, OBOREL fournit des object properties signées comme 'directly *negatively* regulates activity of' qui sont des sous propriétés de propriété causale et aussi de potentiels synonymes de relations causales comme *causally influence* ou *causally related to*. Actuellement, dans OntoBioStat Increase et Decrease ne sont pas des enfants de isCauseof pour des raisons d'inférences simplifiées. Increase et Decrease sont vues comme des spécifications de ces relations causales. Dans OntoBioStat, les synonymes ne sont pas importants. Cependant, ils pourraient être réutilisés en repérant les object properties causales pour éviter de construire un diagramme causal à partir de zéro.

Dans GO-CAM, des relations similaires aux hasMetavariables et à isCauseof sont représentées. Location et Biological Phase (équivalent de Period) sont deux indications spatio

temporelles. Active Entity est reliée à la variable d'intérêt (Molecular Activity) par *enabled by,* object property comparable au principe de relation causale et variable nécessaire.

PHONT est une représentation formelle mais lourde d'un lien causal probabiliste au niveau individuel entre deux variables. Le formalisme d'OntoBioStat, en tant qu'ontologie d'application n'est pas aussi indispensable que pour une ontologie de domaine. De plus, la représentation de la causalité n'est pas individuelle mais populationnelle.

Dans Galton A, 2012, il existe des relations *causal* et *causal-like* et il y a une distinction entre *states*, *event* et *process*. Cette représentation, des états (*states*) qui permettent ou autorisent (object property *allow* ou *enable)* la réalisation d'un processus ou d'un évènement donné est plus formelle que OntoBioStat et a inspiré la création des Necessary_Variable. Dans les diagrammes causaux, cette finesse n'est pas présente non plus. Dans OntoBioStat, que ce soit la relation *allow* , *perpetuate* ou *maintain,* elles sont toutes représentées comme une relation causale standard entre deux variables, l'une d'elles étant une variable nécessaire. Par exemple, la classe "Exist" peut avoir une instance comme "Culture du Tabac" considérée comme une existence nécessaire qui permet la "consommation de tabac".

Dans Ontology-Based Inference for Causal Explanation (Besnard Ph et al., 2008) les auteurs fournissent un cadre de règles pour inférer des relations *explain* à partir de spécification de relations causales et hiérarchique (i.e., is-a). Mais ils ne peuvent inférer que des relations causales direct et indirect et non des corrélations comme le fait de partager un ancêtre commun ou un descendant qui sont primordiales en biostatistiques. Dans l'ontologie sur la radiologie GAMUT, cette même limite d'inférence et de représentation existe avec une relation causale *maycause* transitive sans autres spécifications (Kahn CE, 2016). La classe Path_Modifier est un élément central pour le raisonnement comparée aux précédentes ontologies. Dans un jeu de données, elle permet de prendre compte une variable qui est constante ou qui est sélectionnée dans le modèle statistique afin d'inférer des relations comme partage un descendant ou de bloquer certaines inférences.

Statistical Learning Ontology (SLO) a pour but d'assister l'analyse des données (Behnaz A et al., 2019). Certaines questions de compétences sont similaires à celles de OntoBioStat comme : "What are the variables correlated with ...?". Cependant, les cas d'usage sont sur le marketing digital, la réponse n'est pas basée sur des inférences mais sur des requêtes SPARQL qui vont rechercher de la connaissance déjà spécifiée. SLO ne peut pas répondre aux questions concernant les biais dans la causalité. Un autre point commun

provient du Link Origin (Opinion, Model, Reference) qui se rapproche des object properties d'OntoBioStat cause hypothétique et incertaine.

3.1.2. Epidémiologie

Il existe différentes façon de définir une variable de confusion en langage naturel. Et ces définitions ne peuvent pas être transposées sous la forme de règle sans une adaptation sous peine d'obtenir des inconsistances.. Dans les DAG, c'est un algorithme qui définit si une variable est une variable de confusion. Dans OntoBioStat, la définition correspond à la règle SWRL ou à un axiome ce qui en fait une classe formelle et non plus une définition littérale. Les raisons multiples qui peuvent amener à être un confounder sont donc représentées. Substituant ainsi un algorithme qui peut manquer de complétude et de clarté par de multiples définitions.

De plus, nous avons vu que les biais pouvaient être classés en biais de mesure, biais de sélection et biais de confusion. La distinction en pratique n'a pas d'importance pour l'application de OntoBioStat. En effet, l'intérêt de cette ontologie est la sélection des variables pour diminuer le biais de la relation causale entre exposition et critère de jugement pas de définir le type de biais. Il existe de nombreux sous type de biais avec chaque fois un nom dédié. L'important ici est de savoir si une variable entraîne un biais et si on peut/doit la sélectionner ou si ce biais n'est pas corrigeable. Il a été fait le choix dans OntoBioStat de nommer toutes les variables qui biaisent « confounder » (à quelques exceptions près). Ces variables doivent être considérées pour la sélection.

En épidémiologie ou recherche clinique, la question de recherche causale peut être construite de manière formelle en suivant PICO ou l'Estimand (Lawrance R et al., 2020). L'estimand a cinq attributs : (i) la variable d'intérêt (le critère de jugement ou outcome), (ii) le traitement (l'exposition), (iii) l'évènement intercurrent (arrivant après le début du traitement), (iv) la summary measure (e.g., l'effet sera présenté soit avec un pourcentage soit avec une différence moyenne), et (v) la population. Ce besoin de décrire de manière systématique et homogène la question de recherche est retrouvé dans les métavariables et data properties de OntoBioStat.

3.2. Forces et faiblesses

3.2.1. Constitution du corpus

Concernant les termes extraits à partir des articles sur les DAG, une revue systématique sur les diagrammes causaux et les graphes orientés acycliques a été réalisée par

un seul lecteur. Le corpus a ensuite été revu par deux autres personnes habituées à l'utilisation des DAG ou diagrammes causaux. Les articles de journaux typés comme non épidémiologique ou non biostatistique ont été exclus car considérés comme contenant une information de facto redondante avec les journaux spécialisés. Des articles sur les DAG peuvent être trouvés en dehors de PubMed (e.g., Embase), cependant, les termes concernant l'utilisation des DAG en épidémiologie devraient être exhaustif dans PubMed. Les termes extraits de la littérature : La lecture de ces articles avait un triple intérêt : détecter d'éventuelles nouvelles questions de compétences, élargir la capacité des diagrammes causaux à représenter la connaissance et collecter de nouveaux termes. Les termes en rapport avec les DAGs : le but étant de trouver les termes uniques et synonymes voire leur représentation graphique, les nouveaux concepts étaient toujours définis en début d'article et étaient au nombre maximal de quatre par article. Les reste des termes étaient des termes classiques utilisés couramment (flèche, nœud, confusion, etc), qui de par leur ubiquité n'ont pas fait à proprement parler l'objet d'une extraction. Pour toutes ces raisons, il n'y avait aucun intérêt à utiliser des méthodes de fouille de texte. On ne peut exclure le fait que se cantonner au domaine médical alors que les DAGs sont utilisés dans d'autres domaines a pu entraîner un manque d'exhaustivité.

Seulement quelques guides de reporting du site EQUATOR ont été considérés pour l'extraction des termes. En effet, les guides sont redondants et seuls les plus connus ont été lus. En outre, ceux sur les essais cliniques ont permis d'apporter des détails intéressants sur la définition d'une intervention et pourraient être utilisés pour un développement futur. Cela permettrait d'étendre l'utilisation de OntoBioStat aux essais cliniques randomisés qui peuvent avoir des biais possiblement représentés avec un diagramme causal.
Les guides de bonne pratiques de rédaction d'articles ont été choisis car ils listent un certain nombre d'éléments qu'il est indispensable de voir apparaître dans un article. Cela nous permet d'avancer que l'information devrait être disponible et minimaliste. La vraie difficulté serait de sortir du périmètre en représentation quelque chose d'inutile pour les tâches ou scénarios aux quels l'ontologie est supposée répondre. Par exemple, si j'avais sélectionné plus de guides, il se peut que les idiosyncrasies de certains d'entre eux auraient polluer l'ontologie car rarement utilisés et disponibles. D'autre part, les guides sont très rapidement redondants, aussi, de la même manière que dans une étude qualitative je suis vite arrivé à saturation. Pour finir, il existe de nombreux guides, mais j'ai préféré ceux qui devraient être les plus utilisés de par

leur ancienneté et de par le fait qu'ils ont servi de socle aux guides de la génération suivante ou adaptés à un domaine particulier de recherche.

3.3.2. Développement de l'ontologie

Concernant la validité de l'ontologie, les classes ont toutes une définition concise lorsque cela est nécessaire. Il n'existe pas de classes redondantes. OntoBioStat représente bien toutes les informations pouvant être nécessaires pour la sélection des variables mise à part la forme du lien entre deux variables. Comparée à d'autres ontologies, elle manque de formalisme, cependant le but étant la sélection des variables, elle l'est suffisamment. De plus, un excès de formalisme la rendrait très difficile d'utilisation et lente que ce soit en terme de spécification des instances ou en temps de raisonnement. L'explicabilité des inférences permet aux utilisateurs potentiels de comprendre les résultats. Il serait possible d'étendre l'ontologie sans devoir la refondre complètement. OntoBioStat a été construite en utilisant différentes options de modélisation, que ce soient les axiomes equivalent class, subclass, disjoint class , symetric ou non des object properties et les règles SWRL. Les tests réalisés avec des faux jeux de variables ne renvoient pas d'erreur et permettent de valider les capacités à répondre aux questions de compétence. L'utilisation de l'outil OOPS a permis de mettre en évidence des erreurs (*pit falls*) qui ne sont pas importants dans notre utilisation de l'ontologie. Le raisonneur Pellet permet de s'assurer de la consistance de l'ontologie. Les règles SWRL permettent d'inférer de nombreuses classes et object properties, de plus les inférences sont facilement explicables via l'interface de Protégé. Il n'y a pas de raison particulière à l'utilisation de Pellet plutôt que Hermit ou Fact++.

3.3. Perspectives

OntoBioStat va être testée sur des cas d'usages afin d'illustrer sa capacité à aider dans la construction des diagrammes et proposer des variables à sélectionner, mais aussi expliquer le lien statistique qui existe entre deux variables.

Chapitre 5 : Expérimentations et cas d'usage d'OntoBioStat

Dans ce chapitre, seront présentés les différents cas d'usage sur l'utilisation concrète de l'ontologie OntoBioStat. Les différentes fonctionnalités d'OntoBioStat illustrées dans les cas d'usage sont : (i) la construction du diagramme, (ii) mise en évidence des variables à sélectionner ou patients à exclure, (iii) mise en évidence des biais persistants, (iv) mise en évidence des biais liés aux données manquantes, (v) explication du raisonnement qui permet d'expliquer les résultats au sens large (résultats des inférences et résultats statistiques).

1. Cas d'usages

1.1. Utilisation de OntoBioStat via Protégé

Pour comprendre les différents cas d'usage, il est indispensable d'illustrer l'utilisation de OntoBioStat de bout en bout. Cette premiè partie présentera donc le workflow théorique de OntoBioStat.

Actuellement, l'interface utilisateur correspond à Protégé dans lequel il y a OntoBioStat. Après avoir posé la question de recherche et identifié l'Exposure et l'Outcome, l'utilisateur peut renseigner les instances correspondantes. Par exemple, dans la question de recherche est-ce que le tabac cause le cancer du poumon ?

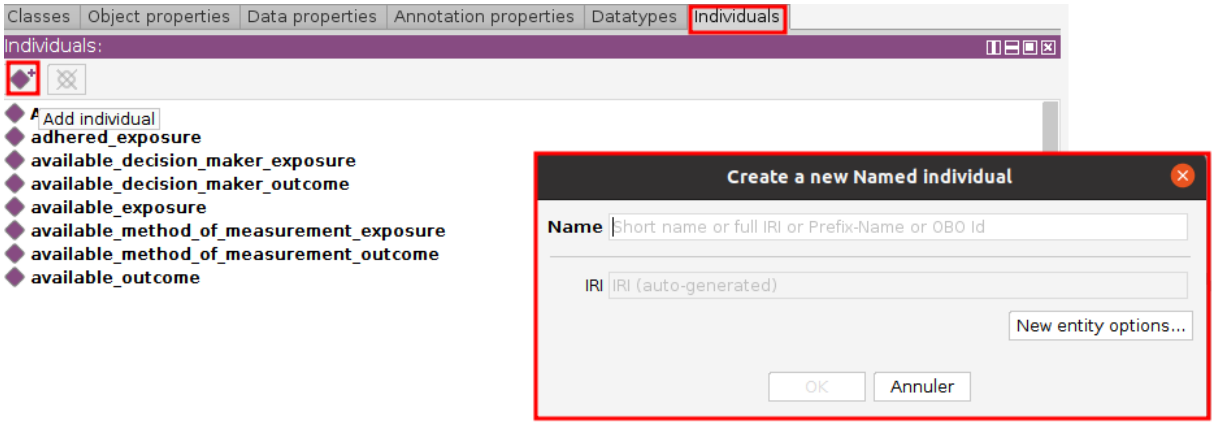


Figure 58. Interface Protégé, ajout d'une instance (variable).

1- L'utilisateur va créer dans Protégé l'instance Cancer du poumon et Tabac (figure 58). Une fois créées, les instances sont classées en Exposure_stressor et Outcome mais aussi dans une classe théorique : la classe théorique Condition pour Cancer du poumon et la classe théorique Health_Behaviour pour le Tabac (figures 59 et 60).

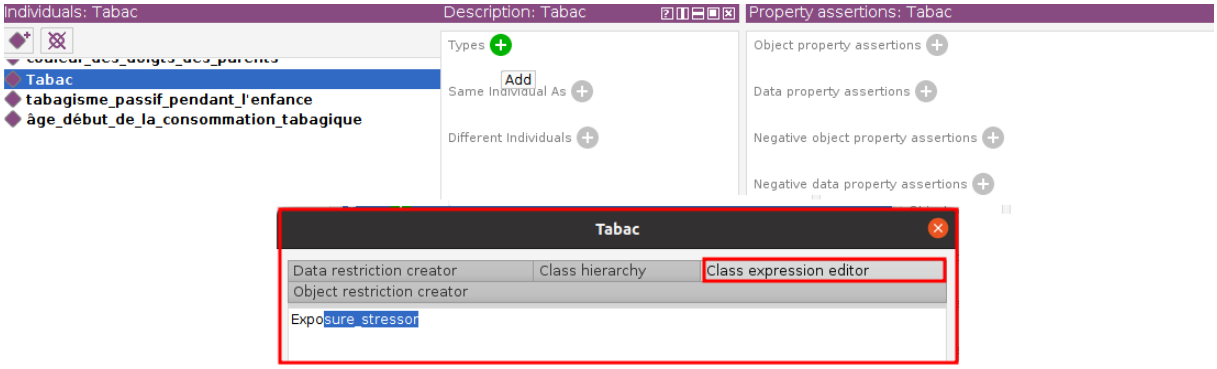


Figure 59. Interface Protégé, classification d'une instance (variable).

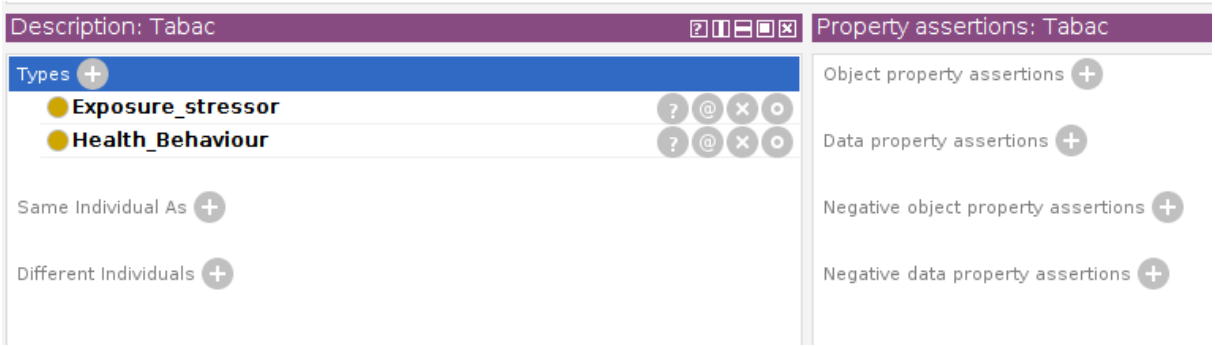


Figure 60. Interface Protégé, instance Tabac classée dans Exposure_stressor et Health_Behaviour.

2- Le raisonneur est activé afin que les instances génériques appartenant à la classe variable nécessaire puissent être connectées par des relations *isCauseof* en fonction des variables théoriques spécifiées juste avant. Elles sont visualisées via les assertions inférées concernant l'Outcome et l'Exposure stressor (figure 61).

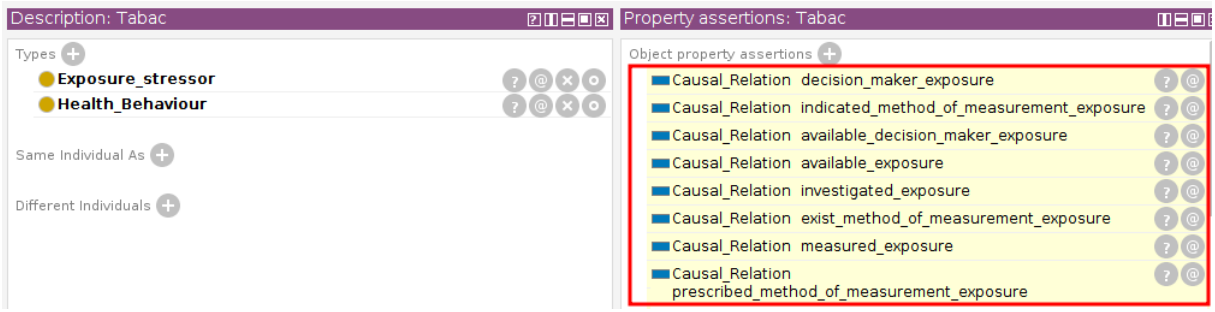


Figure 61. Interface Protégé, visualisation du résultat des inférences: les liens causaux entre le Tabac et les variables nécessaires.

3- L'utilisateur peut ajouter d'autres covariables avec leurs métavariables et data properties correspondantes (figure 62). Ensuite, il pourra connecter les covariables grâce à des object properties *isCauseof*. Par exemple : consommation d'alcool, niveau socio-économique du foyer de naissance, âge de début de la consommation de tabac, tabagisme passif pendant l'enfance, antécédent de cancer de poumon dans la famille, parents fumeurs. Le recueil de la covariable consommation d'alcool s'est fait via un auto questionnaire (Method_of_Measurement), le point de vue est celui du patient (Author_point_of_view), cette variable peut avoir des données manquantes (absence de remplissage non bloquant), a 5 % de données manquantes, et a une raison de données manquantes dont la cause probable est la véritable consommation d'alcool (figure 49). En cas d'erreur, lors des spécifications le raisonneur pourra alerter l'utilisateur et lui rappeler ce qui est attendu comme un case report form (figure 49). Concernant la variable point de vue, il y a autant de points de vue que de patients. Concernant la méthode de mesure, c'est un questionnaire pour tout le monde, elle est constante (Excluded SubClassof Path_Modifier). Il faudra faire attention de bien spécifier que tous les individus sont bien différents. Une seule fois à la fin suffit et cela apparaît dans la section Different Individuals (figure 63).

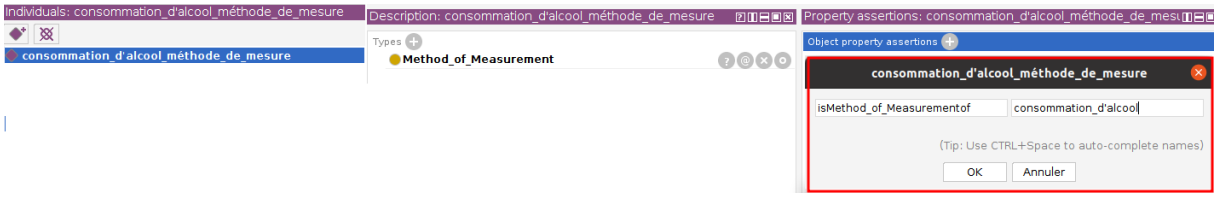


Figure 62. Interface Protégé, spécification de l'object property isMethod_of_Measurementof.

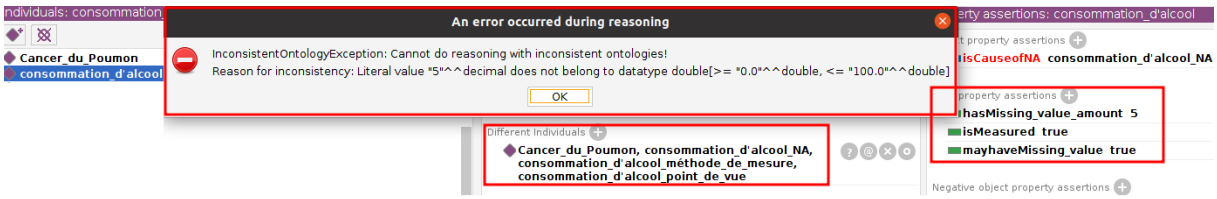


Figure 63. Interface Protégé, inconsistance concernant la spécification de la data property quantité de données manquantes de la variable consommation d'alcool.

4- Le raisonneur est de nouveau activé afin de repérer les inconsistances et les variables proxy potentielles. La covariable parents fumeur n'est pas mesurée mais elle cause une couleur jaunâtre des doigts. Si nous disposions de cette covariable, OntoBioStat nous proposerait de l'utiliser en tant que proxy.

5- Cela signifie remplacer dans le diagramme causal la variable parent fumeur par couleur des doigts (figure 64).

6- Après cela, le raisonneur est activé une dernière fois. Il permet d'obtenir toutes les variables de confusion potentielles. L'utilisateur peut demander pour chaque inférence la raison via 'Explain Inference'(figure 64).

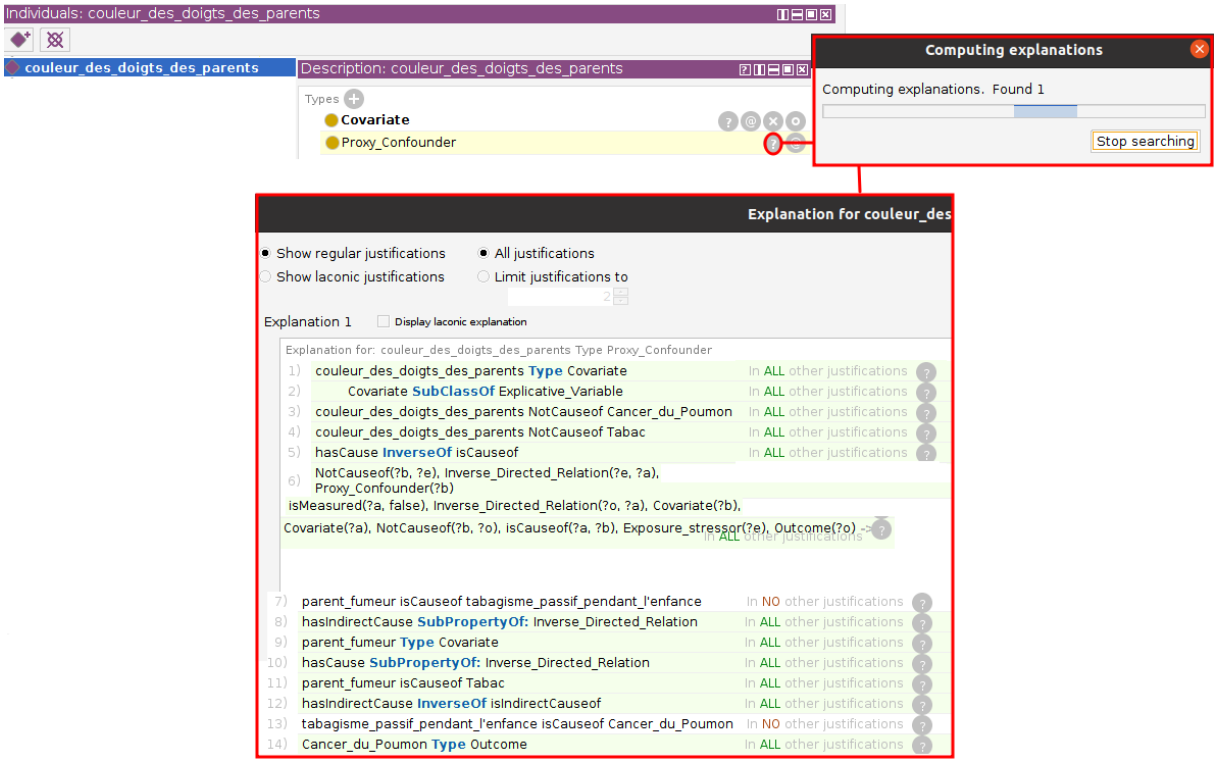


Figure 64. Interface Protégé, inférence de la classe Proxy Confounder et son explication.

Si l'utilisateur désire une visualisation graphique de son graphe, il faut exporter l'ontologie avec ses inférences afin de la rouvrir et d'afficher le graphe via le module OntoGraf de Protégé.

Toutes les étapes nécessaires à l'obtention d'un diagramme causal ontologique grâce à OntoBioStat et l'interface de Protégé sont récapitulées en figure 65.

Figure 65. Figure récapitulative des différentes étapes lors de l'utilisation de OntoBioStat.

(1) l'utilisateur renseigne l'Exposure_stressor et l'Outcome et leur classique théorique respective, (2) le raisonneur est activé et la construction automatique avec les variables nécessaires se produit, (3) l'utilisateur renseigne les autres covariables et les relations causales, (4) le raisonneur est activé, (5) l'utilisateur re travaille le graphe en enlevant et/ou rajoutant des relations, (6) dernières inférences, lecture et explication des résultats.

2. Cas d'usage 1 : construction et analyse du diagramme comparé à un diagramme causal classique

La construction d'un diagramme causal n'est pas régie par de nombreuses règles. Un nœud égale une variable, une flèche égale une relation causale, l'absence de flèche signifie l'absence de relation causale. Il en est de même pour l'interprétation des diagrammes afin de sélectionner les variables, que ce soit en utilisant le back door criteria ou le disjunctive criteria.

Des tutoriels de construction ont été créés afin d'épauler les utilisateurs lors de leur création de diagramme causaux (Digitale JC et al., 2022 ; Vanderweele TJ et al., 2011 ; Stovitz SD, 2019). Ils précisent les fondamentaux de construction et d'analyse décrits plus haut ainsi que l'intérêt de l'utilisation des diagrammes causaux, les conditions d'utilisations. D'autres sont allés plus loin en compilant toutes les « astuces et les pièges » (Suzuki E et al., 2020) comme la représentation des variables lorsque la relation causale est une boucle, lorsqu'il existe des interactions ou des relations signées. Un des tutoriels explique pas à pas comment passer d'une revue de la littérature (i.e., connaissances provenant d'articles) à un diagramme causal (Ferguson KD et al., 2020). Chaque conclusion d'article est compilée dans un 'implied graph' (IG) puis, ces IG sont évalués d'un point de vue causal (temporelle, contrefactuelle, plausibilité et concordante avec la théorie) avant d'être intégrés dans un diagramme final.

Ce cadre assez basique laisse les utilisateurs libres d'adapter leur représentation (e.g., un nœud pour plusieurs variables). Dans une revue sur l'utilisation des DAGs dans les articles publiés avant 2018, un nœud était utilisé pour représenter plusieurs variables dans 17 % d'entre eux. Les auteurs précisaient la présence de variables non mesurées dans seulement 37 % des cas (Tennant PWG et al., 2021).

Un des objectifs de OntoBioStat est la réutilisation totale ou partielle des DAG produits au cours des études. Pour se faire, il est nécessaire de partager le même formalisme lors de la construction et fournir le DAG (seulement 62 % des articles utilisant des DAG le joignent à la publication (Tennant PWG et al., 2021)).

Dans ce cas d'usage basé sur un article scientifique médical, nous verrons : (i) comment passer du contenu d'un article scientifique à une spécification d'instance en suivant le cadre de OntoBioStat ; (ii) les différences saillantes entre diagramme causal ontologique et diagramme causal à partir d'une version inspirée du diagramme construit dans le (i) ; et (iii) les différences entre les inférences de OntoBioStat et celles faites à partir du diagramme causal.

2.1. Exemple d'extraction des informations pertinentes pour la construction d'un diagramme causal

Dans ce cas d'usage le matériel utilisé était un article du BMJ de pharmaco épidémiologie sur l'effet potentiel des incrétinomimétiques comparés aux sulfonylureas sur la survenue du cancer d'un pancréas. L'article est librement accessible sur *https://doi.org/10.1136/bmj.i581*. Cet article a été sélectionné car il présente une étude prototypique observationnelle sur données de vie réelle (base médico-administrative) dans laquelle OntoBioStat pourrait être utilisée. Chaque élément de la méthode a été utilisée pour nourrir l'ontologie. Ces éléments sont classés et traduits manuellement.

Les sources de données sont: administrative and electronic medical record databases (data property **Data_source)**.

Le critère de jugement correspond au cancer du pancréas dans les X années (**Outcome**). La variable théorique correspondante est **Condition**. La méthode de collection du cancer du pancréas consistait à utiliser le codage du diagnostic principal ou associé en CIM-9/10. **Collection_method** Code_CIM-9/10 DP/DAS *hasAuthor_point_of_view* Codeur/Centre. L'exposition correspond à l'incrétin versus sulfonylureas (**Exposure_stressor**). La variable théorique correspondante est **Intervention_Effect**.

Les critères d'inclusion exclusion sont les suivants :

- age (critère d'exclusion: âge <18 ans correspond à data property has**PossibleValue Bounded**),

- duration of medical history (critères d'exclusion: historique médical < 365 jours, correspond à data property has**PossibleValue** Bounded et **Path_Modifier**)

- previous insulin prescription (critère d'exclusion : prise antérieure d'insuline, correspond à data property has**PossibleValue** Constant et **Path_Modifier**) proxy pour diabète avancé (variable is**Measured** FALSE et *isCauseof* prise antérieure d'insuline)

- women with an history of polycystic ovarian syndrome et diagnosis of gestational diabetes in the 365 days before entry to the base cohort (has**PossibleValue** Constant et **Path_Modifier** *isCauseof* prise de Metformine)

La liste des covariables mesurées (is**Measured** TRUE) est la suivante : sex, duration of treated diabetes, duration of follow-up, alcohol related disorders, number of unique antidiabetic drugs received in the 365 days before entry et presence of microvascular complications of diabetes sont deux proxy pour la sévérité du diabète (sévérité du diabète is**Measured** FALSE et *isCauseof* la présence de complication microvasculaire et nombre d'antidiabétique oraux reçus durant l'année écoulée), total number of hospital admissions, total number of unique non-diabetic drugs prescribed, body mass index (has**Type** ordinal), haemoglobin A1c level (has**Type** ordinal), smoking status (has**PossibleValue** Incomplete car valeurs simplifiées en déjà fumé/jamais fumé). Sauf précision contraire, toutes les variables ont le has**Type** dichotomous quand c'est une variable binaire ou le has**Type** quantitative quand c'est une variable quantitative. Les patients avec des données manquantes n'ont pas été exclus. À la place, les variables catégorielles ont une catégorie en plus qui détermine si la valeur était manquante ou non.

Concernant les informations des liens de causalité entre variables : seul les liens causaux avec des proxys potentiels ou des variables d'indication sont décrits. Pour le reste, il faudrait se baser sur la connaissance *a priori* du sujet.

Concernant les informations des données manquantes : absence d'indication du pourcentage de données manquantes par variable, seulement une fourchette concernant l'index de masse corporelle, le niveau d'hémoglobine glyquée et le statut fumeur. Il n'y a aucune information sur le potentiel mécanisme de génération des données manquantes.

2.2. Différence de construction et d'inférence entre diagramme causal et diagramme causal ontologique

Pour illustrer la différence entre diagramme causal (DC) et diagramme causal ontologique (DCO) un plus petit diagramme a été construit en s'inspirant des informations extraites dans le 2.1..

Il inclut les covariables suivantes : comorbidités, effets secondaires des antidiabétique oraux (ADO), co-traitement, diminution de la sécrétion d'insuline. Les variables nécessaires et leur relations causales inférées grâce aux spécifications Outcome is_a **Condition** et Exposure_stressor is_a **Intervention_Effect** : Decision-maker, Investigated, Measured, Indicated, Prescribed. Les relations causales sont spécifiées ci-après et dépendent des covariables, dont les variables nécessaires: (i) les comorbidités (comorbidity) contre indique l'utilisation de l'ADO incretinomimétique et ces comorbidités peut être responsable du développement d'un cancer du pancréas, (ii) le co-traitement (co-treatment) X est souvent prescrit avec l'ADO incrétinomimétique et ce co-traitement est impliqué dans le développement des cancers du pancréas, (iii) les effets secondaires (side effect) des incrétinomimétiques sont suffisamment inconfortables pour pousser les patients à consulter plus souvent et des consultations rapprochées permettent un diagnostic plus précoce du cancer du pancréas, (iv) le cancer du pancréas (pancreatic neoplasm) peut diminuer la production d'insuline et provoquer des hyperglycémies (hyperglycemia), pouvant être pris à tort pour un diabète.

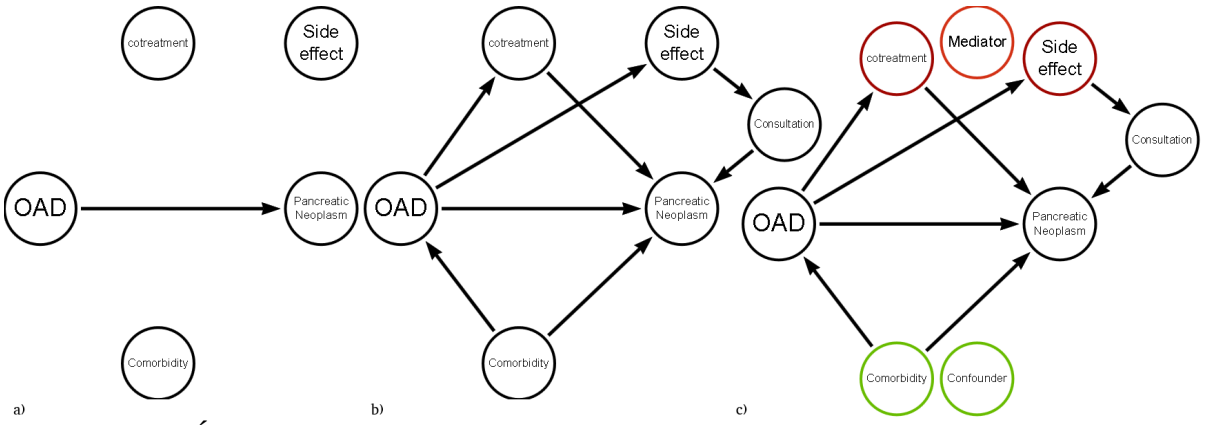


Figure 66. Étapes de construction et d'analyse du diagramme causal: a) un noeud pour chaque variable et la relation causal entre antidiabétique oral et cancer du pancréas; b) relations causales entre les variables; c) mise en évidence des variables de confusion et médiation.

L'analyse du DC pour résoudre la sélection des variables a été basée sur une lecture purement graphique et non l'utilisation d'algorithmes. La causalité inverse implique un graphe cyclique, les diagrammes causaux étant acycliques, la relation de causalité Outcome → Exposure_stressor n'a pu être représentée sur le DC (figure 66).

Concernant le DCO, seule une partie tronquée est présentée par souci de lisibilité (figure 67-71). Quatre étapes de construction sont présentées dans les figures : (a) spécifications de l'utilisateur en rapport avec l'exposition et le critère de jugement ; (b) activer le raisonneur pour la construction automatique des relations nécessaires ; (c) et (d) ajout des covariables et des relations avant de lancer l'inférence finale.

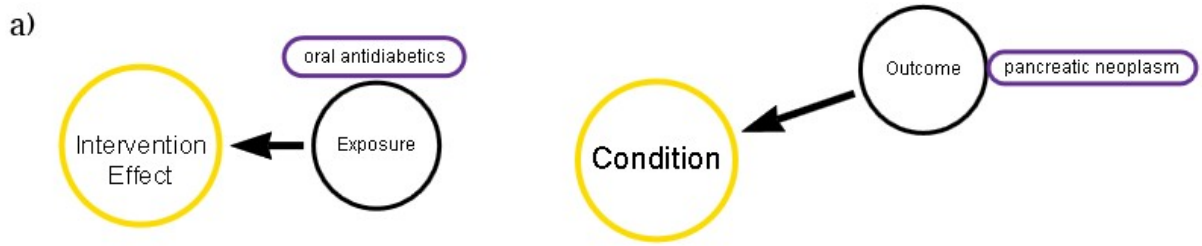


Figure 67. Première étape de construction : spécifier l'outcome, l'exposure et à quelle variable théorique ils appartiennent.

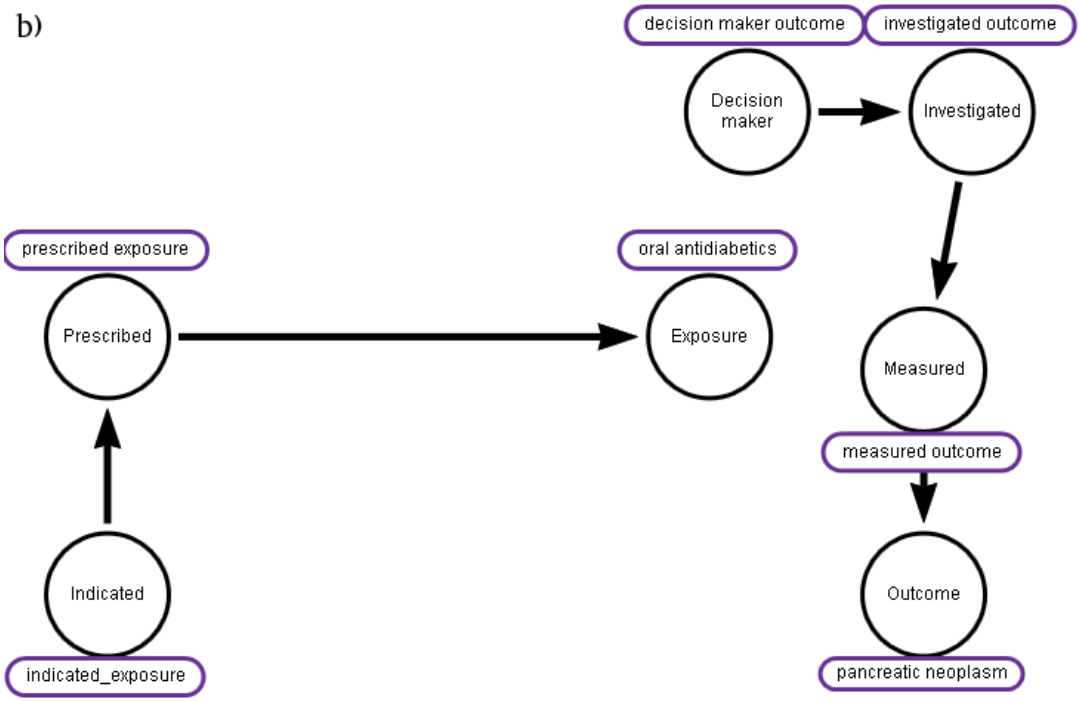


Figure 68. Deuxième étape de construction : inférences qui permet de construire un squelette de variables nécessaires.

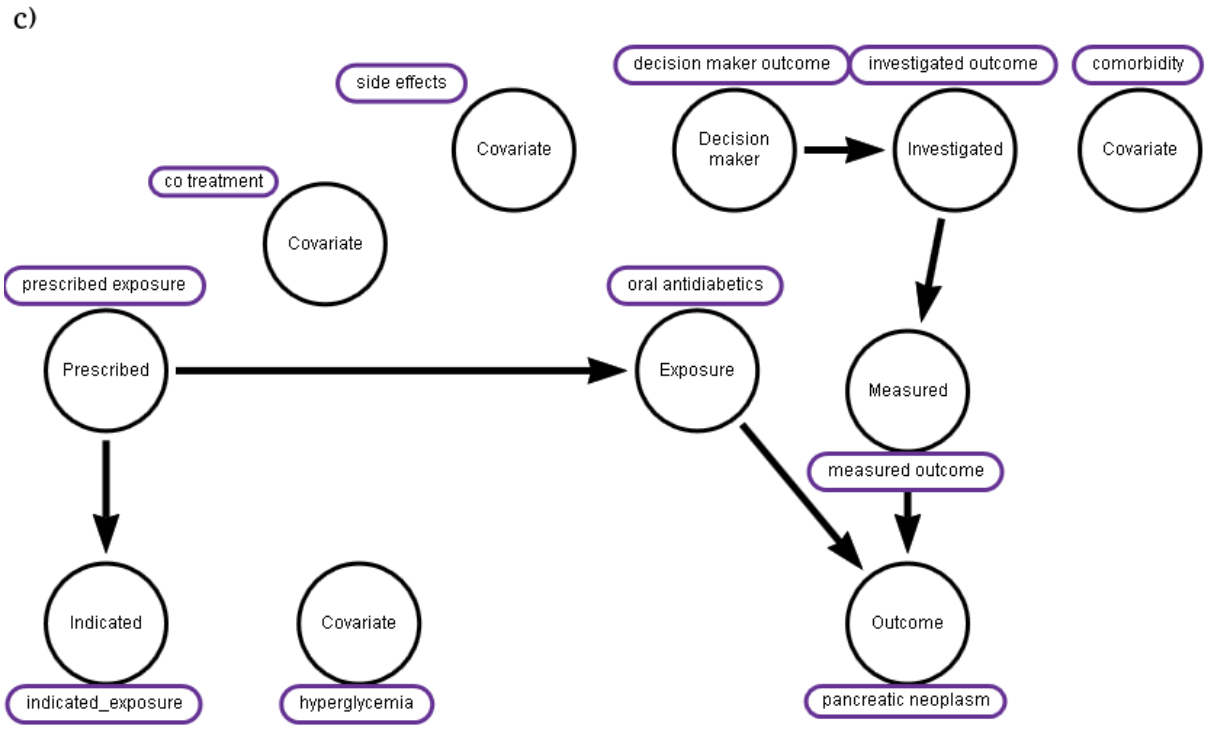


Figure 69. Troisième étape de construction : ajout des covariables.

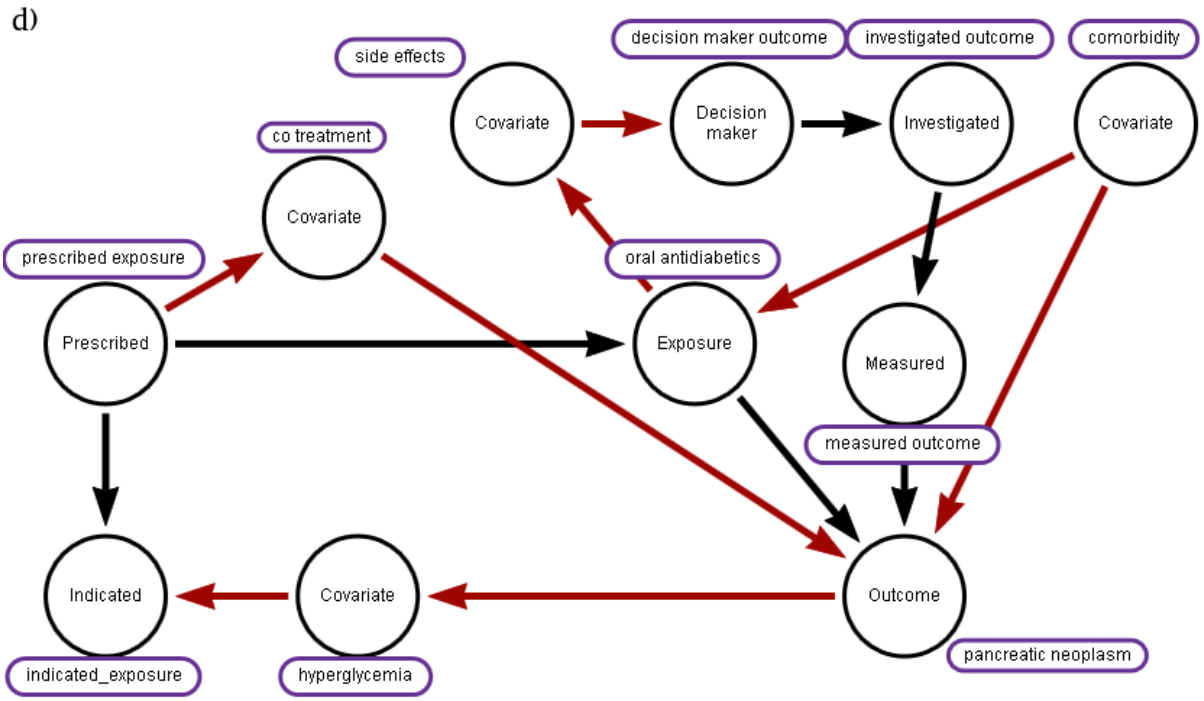


Figure 70. Quatrième étape de construction : création des relations causales entre variables.

Les explications des inférences fournies par Protégé retranscrite en langage naturel donnent : ‘co treatment’ est un **Confounder-like** car ‘co treatment’ *isCauseof* **Outcome** et *hasCause* ‘prescribed_exposure’ qui est un **Indirect_Confounder**; (ii) ‘side effects’ est un **Mediation Differential Confounder** car *hasCause* **Exposure** et *isCauseof* **Necessary_Variable** (1); (iii) ‘comorbidity’ est un **Unadjusted Confounder** car

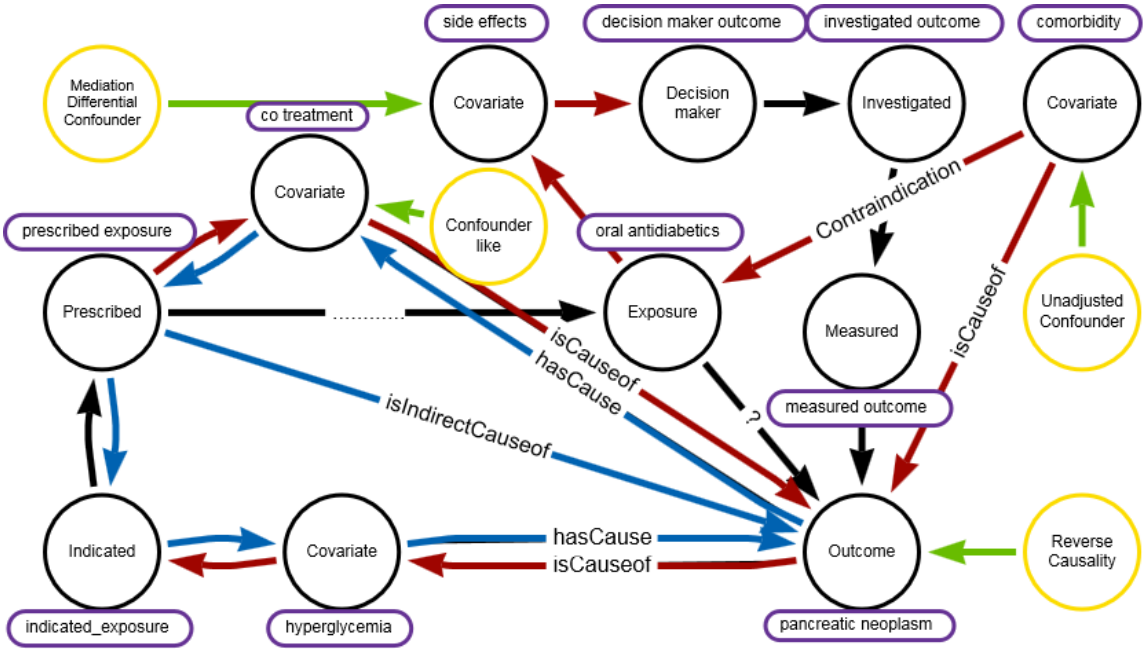


Figure 71: Diagramme Causal Ontologique final simplifié après inférences (en jaune).

Contraindication de **Exposure** et *isCauseof* **Outcome**; (iv) **Outcome** est **Reverse_Causality** car **Outcome** *isIndirectCauseof* **Exposure**.

Sans les variables nécessaires, ‘co treatment’ et ‘side effects’ sont vues comme des covariables qui ne biaisent pas le vrai effet causal mais comme variables sur lesquelles il ne faut pas ajuster (médiation) sous peine d’augmenter le biais. Même avec les variables nécessaires, ‘side effects’ requiert un raisonnement adéquat pour être considéré comme un candidat potentiel pour la sélection. Sans la spécification d’une cause suffisante, la covariable comorbidité est vue comme un candidat potentiel pour l’ajustement alors que les patients devraient être exclus. Pour finir, la causalité inverse a pu être représentée sur le diagramme causal ontologique puisque ces inférences adaptées permettent de faire des exceptions concernant le côté cyclique de la représentation (Tableau 7).

Tableau 7: Conclusions obtenues en fonction de la représentation de la connaissance utilisée

COVARIABLE	OntoBioStat	Diagramme Causal	Pourquoi
Co treatment	Sélectionner	EXCLURE	Necessary variable
Side effect	Sélectionner	EXCLURE	Necessary variable + raisonnement adapté
Comorbidity	EXCLUSION (patients)	Sélectionner	Sufficient cause + raisonnement adapté

2.3. Discussion

Peupler OntoBioStat à partir d'un article est facilitée par le fait que les auteurs suivent les guides de bonne pratique de reporting et que OntoBioStat a été construite en partie grâce à ces guides de bonne pratique. Les relations causales entre les variables ne sont pas toutes spécifiées dans un article scientifique, il n'est donc pas possible de construire un diagramme causal complet sans partir d'une connaissance médicale théorique et pratique dont disposent les cliniciens. Un diagramme causal ontologique peut contenir toutes les informations disponibles dans un article ainsi que les variables nécessaires ce qui permet une construction systématique.

En contrepartie, la construction du diagramme demande plus d'attention et est plus laborieuse qu'un diagramme causal car le nombre d'éléments à spécifier est plus important. Les inférences possibles sont supérieures à celles utilisées en pratique telles que le *disjunctive criterion* ou le *backdoor criterion* (VanderWeele TJ et al., 2011 ; Pearl J 2001). En effet, il est possible de représenter la causalité inverse, les patients à exclure lorsqu'il existe des contre indication et considérer certaines variables de médiation comme des variables de confusion.

Dans les cas d'usages suivants, nous verrons comment OntoBioStat peut aider à interpréter les résultats statistiques et comment OntoBioStat peut aider à sélectionner les variables dans une vraie étude rétrospective.

3. Cas d'usage 2 : Interpréter les résultats statistiques

Interpréter les associations statistiques retrouvées dans les études observationnelles est une discussion ouverte dans la communauté des épidémiologistes. Par exemple, lorsqu'un risque relatif = 2 IC95 % [1,5;2,3] est retrouvé, plusieurs façons de rapporter le résultat peuvent être retrouvées : certains diront que c'est un effet causal, une association, un prédicteur, un proxy, un facteur de risque ou encore tout autre chose. Chacun de ces termes a une portée différente et la façon de rapporter les résultats ne doit pas être prise à la légère, au risque de fournir une mauvaise interprétation. Ils peuvent être expliqués de manière précise si

nous connaissons la structure causale existante entre les variables. Les raisons pour lesquelles deux variables sont corrélées sont les suivantes : A cause B ou B cause A, A et B ont une cause commune, ou A et B ont une conséquence commune constante. Les notions spatiales et temporelles sont aussi importantes car deux évènements peuvent être corrélés car ils se déroulent au même endroit. Les chercheurs peuvent être découragés d'expliquer toutes les associations fausses ou pas, inattendues ou évidentes, retrouvées pendant le processus de la recherche. En effet, le nombre de variables et de tests statistiques vont croissants. De plus, les plus jeunes pourraient mal interpréter les associations statistiques issues des tests bivariés ou des modèles multivariés (Westreich D et al., 2013).

Dans ce cas d'usage, OntoBioStat tentera d'expliquer les associations statistiques retrouvées entre variables, ce qui correspond à la 7ème question de compétence.

3.1. Matériel et Méthode

Afin d'illustrer la faculté d'OntoBioStat à répondre à la question « Quelle type de relation existe entre deux variables ?» et expliquer cette relation, ce cas d'usage est basé sur un vrai jeu de données dérivés de celui du cas d'usage suivant sur la chirurgie des diverticulites perforées avec péritonite. Les données, utilisées dans ce cas d'usage, sont issues d'un recueil rétrospectif sur dossier patient pour la réalisation d'une étude observationnelle monocentrique réalisée au CHU de Rouen. Un total de 16 variables et 102 observations sans données manquantes ont été utilisées. Les 16 variables étaient les suivantes : sexe, chirurgie de nuit, indice de masse corporelle, passage en réanimation, score Hinchey, expérience du chirurgien, diabète, type de procédure, spécialité du chirurgien, Clavien-Dindo, age, ASA-Score, arthritic disease, cancer, traitement par corticostéroïde, immunodépression.

Le nombre d'observations du jeu de données a été augmenté de 102 à 300 en utilisant une méthode de tirage au sort avec remplacement dans le but d'obtenir une puissance statistique suffisante et être capable d'observer des relations causales plus indirectes (plus faibles).

L'association entre chaque variable a été mesurée avec un test de corrélation de Pearson. L'association entre deux variables a été exprimée avec une p-valeur (pas d'estimation du rho de Pearson ni d'intervalle de confiance).

Chaque association statistiquement significative (p-valeur <0,05) a été expliquée par les inférences de OntoBioStat. L'interface de Protégé fournit une option 'explique les inférences' qui permet de voir les différentes étapes du raisonneur qui l'ont mené à une

inférence donnée. Les object properties ne sont pas mutuellement exclusives, on s'attend à ce que deux relations par exemple, *Share_ancestor* et *isCauseof* soient deux explications possibles et/ou conjointes d'une relation statistique retrouvée.

Pour démontrer l'impact de l'utilisation de la classe Path_Modifier, deux modèles multivariés ont été calculés. Les résultats des modèles ont été exprimés avec l'estimateur (rapport de côtes) et la p-valeur avant et après l'ajustement sur une variable.

Les utilisateurs, par l'intermédiaire de l'éditeur d'ontologies Protégé, ont renseigné les informations suivantes basées sur la connaissance médicale : (i) le nom des instances qui sont des **Covariables** (mesurées ou non) ou des **Path_Modifier** ; ensuite (ii) les liens de causalité entre les variables (uniquement *isCauseof*), comme si un diagramme causal ou un graphe orienté acyclique était dessiné ; finalement (iii) le raisonneur Pellet a été activé. Les cinq règles SWRL suivantes ont été utilisées ((1),(2),(3),(4),(5)) et les inférences étaient donc circonscrites aux object properties représentées dans la figure 72.

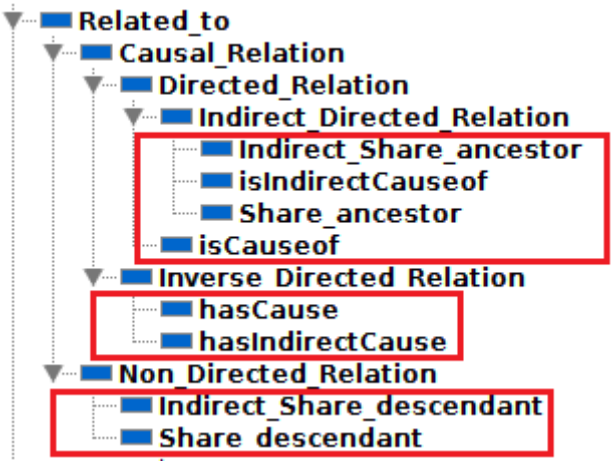


Figure 72: Object properties inférées

Dans ce cas d'usage, toutes les options d'OntoBioStat n'ont pas été utilisées comme les Necessary_Variable ou les Theoretical_Variable ou encore les object properties Contre_indication ou Interact_with.

3.2. Résultats

Un total de 28 différentes instances ont été créées, soit 16 variables mesurées et 12 non mesurées. Quarante-huit relations *isCauseof* ont été spécifiées. La représentation graphique fournie par Protégé permet d'analyser grossièrement les relations entre variables, mais le nombre de 28 variables rend déjà la tâche très laborieuse et difficile (Figure 73). L'activation

du raisonneur Pellet a permis d'inférer 1.939 object properties en moins de 5 secondes (Figure 74).

Un total de 48 p-valeurs étaient significatives (Figure 75) et 31 étaient expliquées par les object properties suivants : (i) dix *isCauseof* (e.g., corticosteroid therapy *isCauseof* immunosuppression); (ii) 12 *isIndirectCauseof* (e.g., sex *isIndirectCauseof* ASA score parce que sex *isCauseof* arthritic disease, arthritic disease *isCauseof* organe failure at the beginning *isCauseof* ASA score) ; (iii) dix *Share_ancestor* (e.g., bmi
et diabete *Share_ancestor* health behavior) ; (iv) 11 *Indirect_Share_ancestor* (e.g., surgical complication at 90 days (Clavien-Dindo) *Indirect_Share_ancestor* avec post operative intensive care unit car Hinchey *isCauseof* organe failure at the beginning et surgical complication at 90 days (Clavien-Dindo) et organe failure at the beginning *isCauseof* post operative intensive care unit). Deux explications (deux object properties différents) ont été fournies pour 12 associations.

Figure 73. 28 instances de covariable et 48 relations isCauseof.

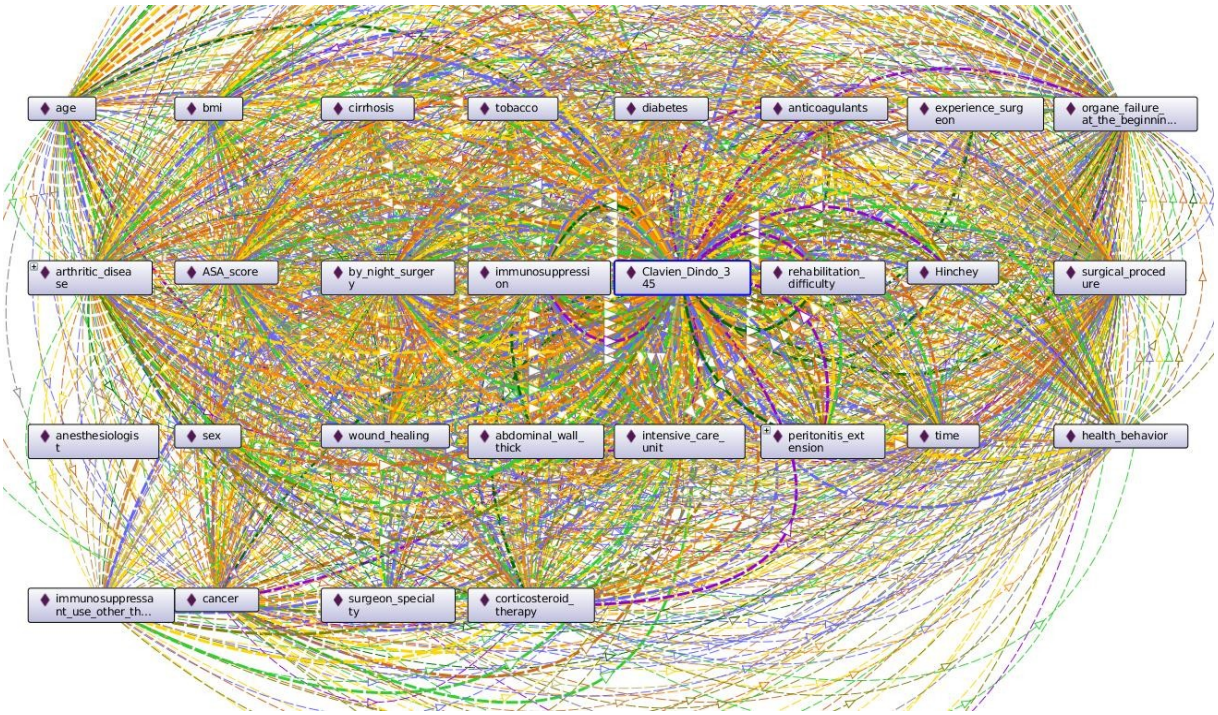


Figure 74. Résultats des inférences.

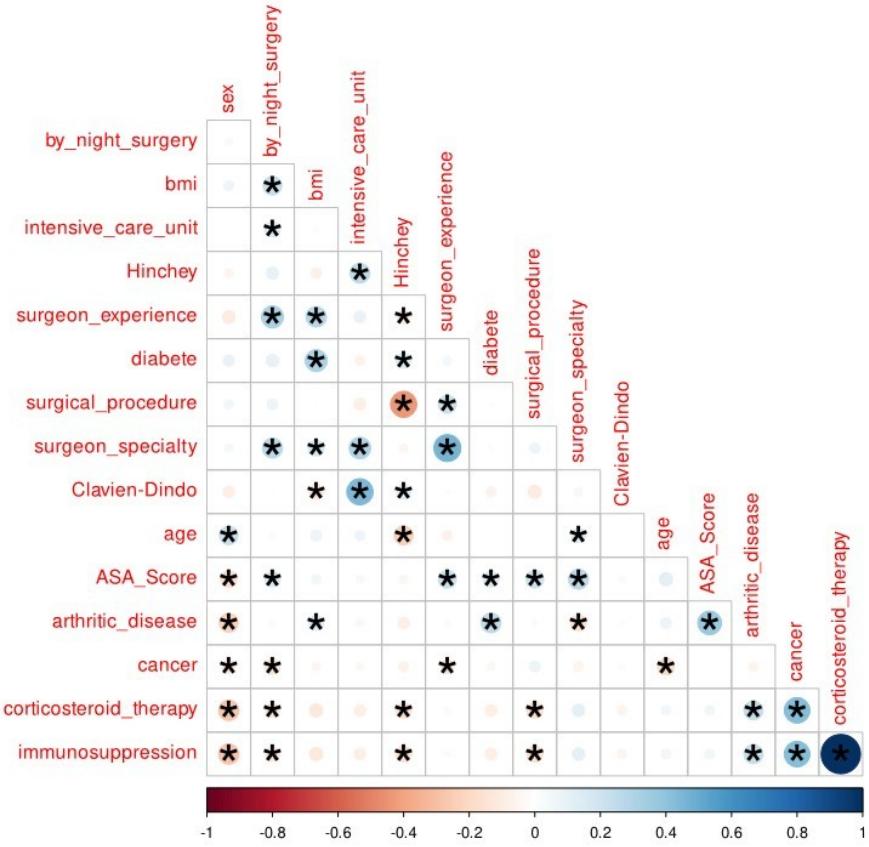


*Figure 75. Matrice de corrélation, *: statistiquement significatif.*

Le premier modèle multivarié est une régression linéaire avec l'ASA score en tant que Outcome et sex et arthritic disease en tant que variables explicatives. Avant l'ajustement la variable sex et l'ASA score sont associées parce que sex *isIndirectCauseof* ASA score (p-value =0,004, estimate = -0,25) (figure 76). Après ajustement sur arthritic disease, la variable sex n'est plus associée avec ASA score parce que le chemin a été intercepté par le 'Path_Modifier'arthritic disease (p-value = 0,09, estimate = -0,14) figure 78.

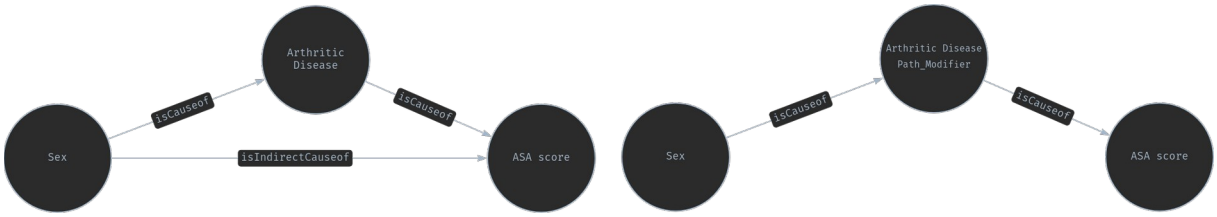


Figure 76. Relations causales avant et après ajustement sur arthritic disease.

Le second modèle statistique multivarié est une régression logistique avec immunosuppression en tant que Outcome et surgeon specialty et by night surgery comme variables explicatives. Avant ajustement, surgeon specialty et immunosuppression ne sont pas associés (p-value = 0,319, estimate = 0,72). Après ajustement sur by night surgery, la variable immunosuppression est presque statistiquement associée avec surgeon specialty car ils partagent un descendant commun *Share_descendant* (p-value = 0,08, estimate = 1,36). Dans l'explication donnée figure 79, il y a deux règles SWRL, 5 inférences provenant des relations hiérarchiques ou inverses et 8 spécifications.

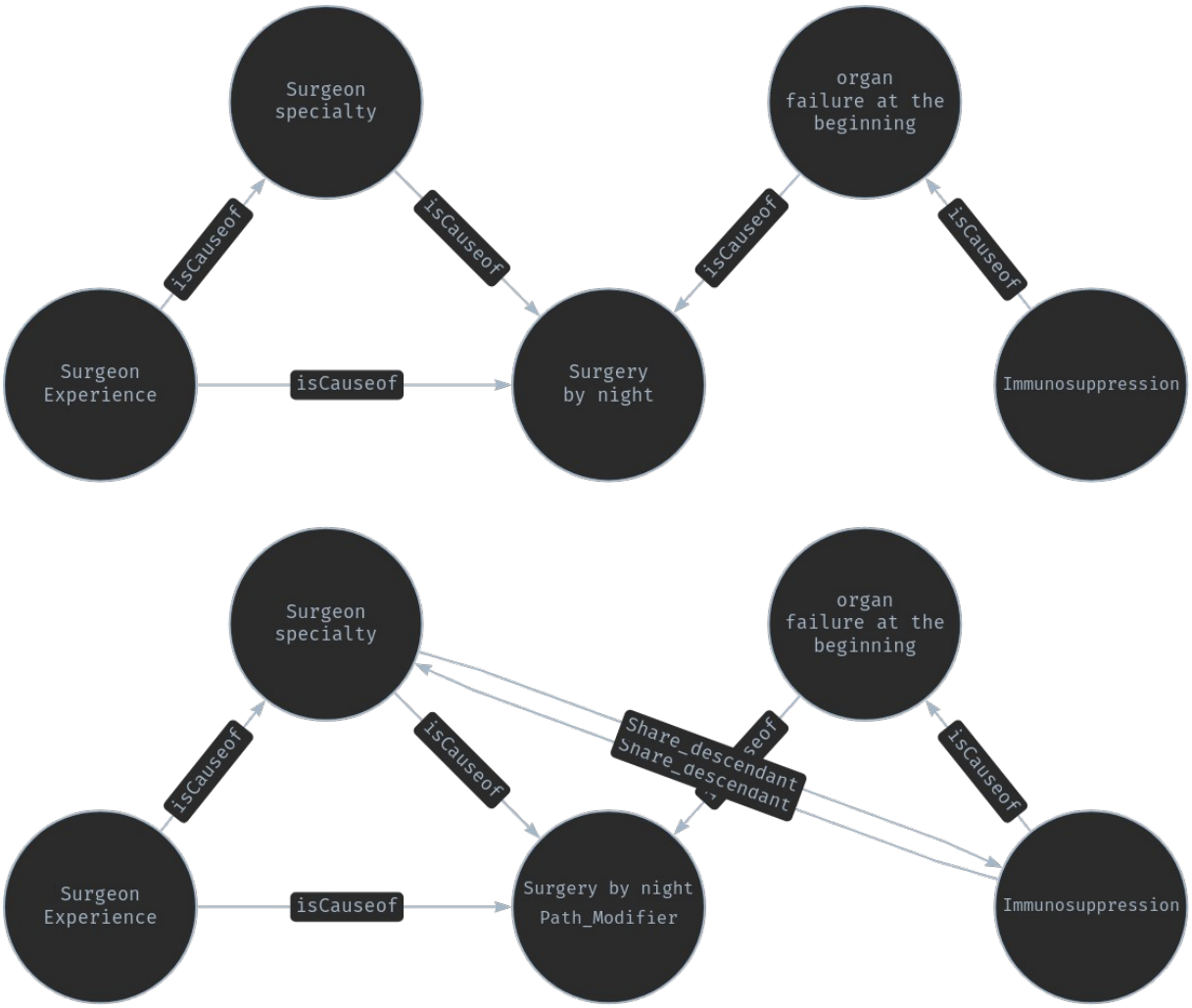


Figure 77. Relations causales avant après ajustement sur by night surgery.

Explanation for: sex isIndirectCauseof ASA_score

Covariate **SubClassOf** Inferred_Variable
hasCause **InverseOf** isCauseof
Inverse_Directed_Relation(?b, ?a), Inverse_Directed_Relation(?c, ?b), Inferred_Variable(?a), Inferred_Variable(?b) -> isIndirectCauseof(?a, ?c)
sex isCauseof arthritic_disease
organe_failure_at_the_beginning isCauseof ASA_score
hasIndirectCause **SubPropertyOf:** Inverse_Directed_Relation
organe_failure_at_the_beginning **Type** Covariate
arthritic_disease **Type** Covariate
hasCause **SubPropertyOf:** Inverse_Directed_Relation
arthritic_disease isCauseof organe_failure_at_the_beginning
hasIndirectCause **InverseOf** isIndirectCauseof
sex **Type** Covariate

Figure 78. Explication de la relation statistique entre sex et ASA score.

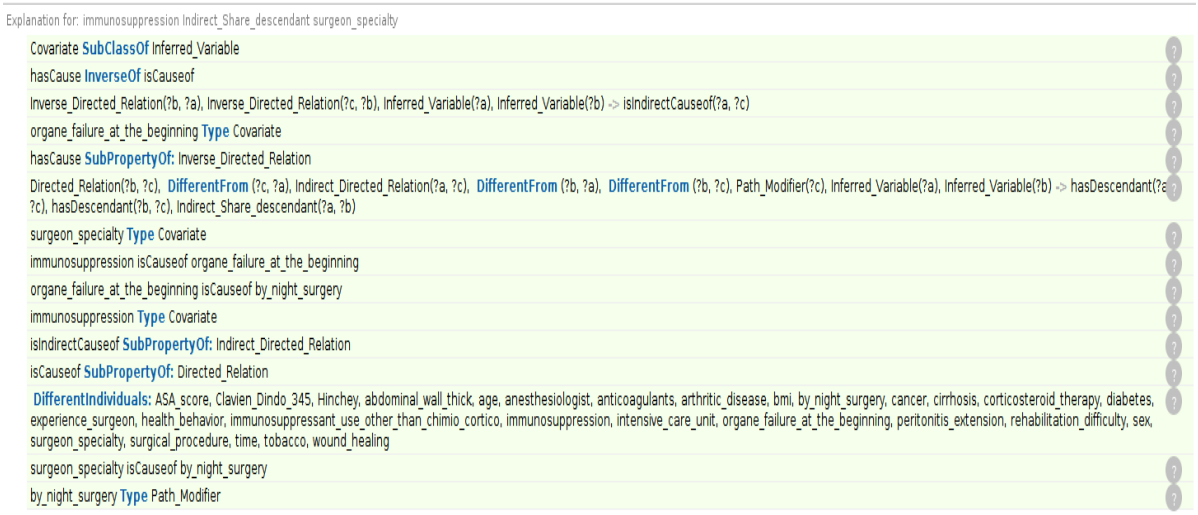


Figure 79. Explication de la relation statistique entre Immunosuppression et surgeon specialty après ajustement sur by night surgery.

3.3. Discussion

Dans ce cas d'usage OntoBioStat a été utilisée pour expliquer les associations statistiquement significatives entre 16 variables. OntoBioStat a réussi dans 64,5 % des cas, 35,5 % restant inexpliquées. Ces résultats peuvent être expliqués par : (i) des erreurs de représentation de la connaissance initiale, (ii) des tests statistiques multiples (en effet, avec 120 tests statistiques près de 6 résultats significatifs le sont par erreur, de plus la non indépendance des variables peut avoir un effet en chaîne), (iii) le jeu de données a été artificiellement augmenté, (iv) de plus, une p-valeur<0,05 ne veut pas nécessairement dire que l'association est cliniquement significative.

Ces premiers résultats, permettent aussi de revenir sur la construction du diagramme pour vérifier s'il n'y a pas eu des erreurs. Ici, une des relations qui n'était pas inférée concerne l'âge et le sexe. De prime abord, il n'y a aucune relation causale entre âge et sexe. Ici nous sommes dans une maladie (diverticulite perforée avec péritonite) qui concerne principalement les personnes âgées avec un régime pauvre en fibre, donc les hommes âgés. Si nous tenions compte de cet élément (sex *isCauseof* diverticulite et age *isCauseof* diverticulite et diverticulite est un **Path_Modifier**), la relation statistique trouvée entre âge et sexe correspond à *Share_descendant*. Ce qui expliquerait aussi la relation sexe et cancer.

OntoBioStat fournit une explication claire là où un sempiternel « association statistique » aurait pu servir d'explication. En effet, il est possible de justifier « l'apparition » ou la « disparition » d'une p-valeur significative grâce aux object properties. Il est possible de

remonter le raisonnement. Le nombre total d'object properties inférés est très important (n=1 939) mais pas toujours utiles pour un biostatisticien voire redondants. En effet, la spécification de *isCauseof* produit les cinq inférences suivantes à cause de la hiérarchie et des relations inverses : hasCause, Inverse_Directed_Relation, Directed_Relation, Causal_Relation, and Related_to. Ces cinq-là ne sont pas toujours utiles mais sont nécessaires pour simplifier les règles SWRL et délivrer une information plus superficielle. Dans une interface finalisée, les inférences devraient être fournies sans informations inutiles et se concentrer par défaut sur les huit object properties encadrées en rouge de la figure 58.

Contrairement aux autres approches manuelles ou visuelles utilisant les DAG ou les diagrammes causaux, OntoBioStat fournit des inférences automatiques avec des explications. Les approches visuelles devenant inefficaces voire impossibles avec un nombre croissant de variables.

En conclusion, malgré cette option d'explication, il est important de noter que OntoBioStat ne prédit pas l'apparition d'une association statistique (il faudrait connaître la puissance de l'étude et la force d'association *a priori*), ni leur signe, ni leur force.

4. Cas d'usage 3 : Mise en évidence des variables de confusion dans une étude réelle

Aujourd'hui, les outils ou packages R pour la sélection des variables à partir de graphes orientés acycliques ou diagrammes causaux sont rares (ggdag : Barrett M, 2023 , dag-R : Lutz P Breitling et al. 2021), Dagitty : Textor J et al., 2017 ; Ankan A et al., 2021). Les auteurs de Dagitty ont aussi créé une application web qui permet de tracer un diagramme à la main, d'exporter le code produit et d'obtenir le jeu minimal de variables (figure 80).

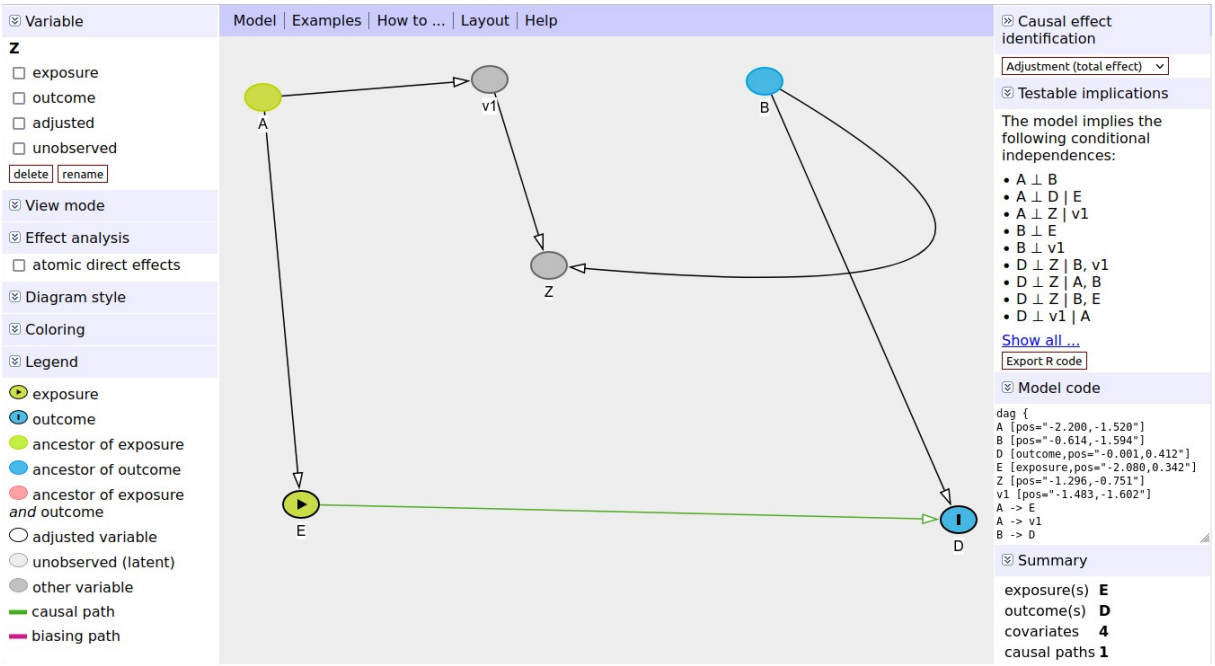


Figure 80. Figure de l'interface web de Dagitty.

Concernant la construction du diagramme avec l'interface web de Dagitty :

Volet de gauche : il est possible de spécifier si la variable **Z** est l'outcome, l'exposition, si c'est une variable ajustée et si elle est observée ou non (mesurée). La légende permet de voir les nœuds outcome, exposition et les covariables (other variable), celles qui sont des ancêtres de l'outcome, de l'exposition ou des deux, les variables latentes ou non observées (non mesurées), le lien causal étudié (causal path) (Tableau 8).

Volet de droite : il est possible d'obtenir toutes les indépendances conditionnelles. A ⊥ B veut dire A est indépendant de B, autrement dit A est non corrélé statistiquement à B. A ⊥ D| E veut dire A est indépendant de D conditionnellement à l'ajustement sur E, c'est à dire que A et D sont corrélés sauf en cas d'ajustement sur E. En dessous, il est possible d'exporter le code pour l'utiliser avec le package du même nom. Et pour finir, il y a un petit résumé : nombre de covariable, nom de l'exposition, nom de l'outcome.

Au centre : il est possible de tracer le diagramme causal en utilisant des nœuds et des flèches unidirectionnelles ou bidirectionnelles.

Dans Dagitty, pour connaître le jeu minimal de variables qu'il faut sélectionner, c'est l'indépendance conditionnelle entre E et D qu'il faudra regarder. E ⊥ D | X X X (les X correspondent au jeu minimal de variables).

Tableau 8: Termes utilisés dans Dagitty et leur équivalent dans OntoBioStat

Dagitty	Exposure	Outcome	Other variable	Ancestor of exposure and outcome	Ancestor of outcome	Ancestor of exposure	Adjusted variable	Unobserved
OntoBioStat	Exposure_stressor	Outcome	Covariate	Direct Confounder et Indirect Confounder	IsCauseof Outcome ou IsIndirectCauseof Outcome	Idem avec Exposure_stressor	Path_Modifier	isMeasured = FALSE

Dans ce cas d'usage, nous verrons comment OntoBioStat peut être utilisée en condition réelle comme Dagitty et non pas à partir d'un article. Ce cas d'usage s'appuie sur le travail réalisé dans le cas d'usage précédent. Les différences entre Dagitty et OntoBioStat seront discutées.

4.1. Matériel et Méthode

Une étude observationnelle rétrospective multicentrique (**Location**) sur la chirurgie digestive a été utilisée pour illustrer l'utilité de OntoBioStat pour la sélection des variables.

La question de recherche était la suivante : "Est-ce que la procédure chirurgicale de résection et anastomose protégé (RAP), comparée à la résection anastomose non protégé (**Exposure_stressor** et **Intervention_effect**), réduit le nombre de complications chirurgicales graves (**Outcome** et **Condition**) à 90 jours ?" L'**Exposure_stressor** has**Definition** comme une stomie protectrice, implique la mise en place d'une stomie pour prévenir les fuites anastomotiques, évitant ainsi la péritonite, le sepsis voire le choc. L'**Outcome** Clavien-Dindo >= IIIb est défini comme une révision chirurgicale ou une complication mettant la vie en danger nécessitant des soins intensifs dans les 90 jours post-opératoires, y compris les complications suivant le retrait de la stomie, excluant les complications immédiates post-opératoires.

Les patients sélectionnés ont eu une péritonite d'origine diverticulaire (**Path_Modifier**), causée par des facteurs tels que le sexe et l'âge (âge et sexe *isCauseof* péritonite diverticulaire), avec un Hinchey entre 3 et 4 (péritonite purulente ou fécale) has **PossibleValue** Bounded et ont été traités soit par résection anastomose protégée (RAP), soit par résection anastomose non protégée (RANP). Habituellement, il existe un troisième traitement appelé le Hartman, cependant les patients ayant subit cette opération ont été exclus. La collecte de données a été effectuée manuellement par un acteur désigné dans chaque centre, à partir de rapports médicaux et d'observations (**Collection_method** "manuel humain", has**Data_source** "Electronic Health Report").

Les variables (**Covariable** sauf mention contraire) mesurées (is**Measured** TRUE) :

- MetaVariable **Period** chirurgie de nuit.
- MetaVariable **Decision-maker** expérience du chirurgien (sénior, junior), spécialité du chirurgien (généraliste, spécialiste) peuvent impacter le choix de la technique.
- MetaVariable **Location** Lieu de la chirurgie (la localisation de la structure de soin) peut influencer le choix de la technique et le protocole post-operatoire lui même impactant l'**Outcome** (Center *Cause_Hypothetically* protocole post-opératoire).
- ASA-Score (has**Missing_value_amount** 2%), un indicateur des comorbidités du patient ainsi que de son état avant la chirurgie (éventuel choc, etc). *NotCauseof* **Outcome**. Sa mesure dépend de **Collection_Method** (il n'y avait pas d'instruction spécifique si l'ASA-Score devait être recalculé, imputé ou non) et de **Author_point_of_view** (anesthesiologist). Les données manquantes pouvaient être causées par la vraie valeur de l'ASA-Score (*isCauseofNA* ASA-Score_NA)
- Indice de masse corporelle (variable avec données manquantes) indicateur du comportement de santé nutritionnel et de la sédentarité impliquée dans les maladies artérielles et la circonférence abdominale.
- Diabète. Idéalement, la variable diabète devrait avoir plusieurs valeurs, notamment le type 1 et le type 2, avec ou sans traitement à l'insuline. Cependant, étant donné que nous disposons uniquement d'une représentation binaire du diabète (oui ou non), cette variable has**Possible_Value** Incomplete. De plus, étant donné que l'absence de diabète n'est pas toujours explicitement documentée dans les dossiers médicaux, l'absence totale de toute mention de diabète a été considérée comme une réelle absence de la maladie. Cela a été traité grâce à une imputation native avec une valeur de 0, d'où la mention.
- Âge est une variable d'indication d'une procédure plutôt qu'une autre, et un âge avancé implique des complications post-opératoires plus fréquentes.
- Sexe n'est pas relié à l'**Exposure_stressor** ni l'**Outcome**.
- Passage en réanimation à la sortie du bloc opératoire (has**Missing_value_amount** 10%). Ce passage en réanimation n'est pas causé par un type de chirurgie plutôt qu'un autre et n'est pas responsable d'un Clavien-Dindo ≥ 3b (*NotCauseof* **Outcome**).
- L'immunosuppression, qui peut résulter des traitements tels que les corticostéroïdes ou la chimiothérapie ainsi que de la maladie sous-jacente, est associée (*Cause_Under_Validation*) à l'apparition de collections abdominales post-opératoires et à un risque accru de mortalité post-opératoire (has**Possible_Value** Incomplete. mayhave**Missing_value** FALSE).

Étant donné que les patients à haut risque de complications ou avec une instabilité hémodynamique devraient être exclus, le choix entre RAP et RANP proviendrait essentiellement de la méta variable **Decision-maker** qui sera ici l'expérience du chirurgien. Les comorbidités du patient ainsi que l'état au moment de la chirurgie pourraient servir au chirurgien dans la décision de son traitement mais pourraient aussi impacter les risques de complications ultérieures. La guérison des tissus dépend de la santé du réseau artériel sous-jacent, l'épaisseur de la paroi peut rendre difficile le port d'une stomie, l'immunodépression est associée à une collection abdominale plus fréquente et une mortalité augmentée.

Les variables non mesurées ou mal mesurées (is**Measured** FALSE) :

- l'instabilité hémodynamique ou défaillance d'organe initiale n'était pas mesurée. Cependant, dans notre sous groupe, elle devrait être constante (has**PossibleValue** constant) car l'instabilité hémodynamique est une contre indication à l'anastomose (l'instabilité hémodynamique *Contraindication* RAP & RANP). Après discussion avec les chirurgiens, il n'est pas possible de confirmer cette hypothèse en l'absence de données. Dans un faible pourcentage de cas, selon le degré d'instabilité, une RA aurait été pratiquée à la place d'un Hartman. L'object property signée *Contraindication* ne sera donc pas utilisée dans notre cas d'usage, même si elle pourrait faire l'objet d'une analyse de sensibilité. L'ASA score à partir du niveau 4 pourrait correspondre à un état de défaillance (défaillance *isCauseof* ASA). Un passage en réanimation à la sortie du bloc opératoire est une conséquence d'un état de défaillance existant ou latent au début de la chirurgie (défaillance ou risque de défaillance *isCauseof* passage en réanimation).

Toutes les variables non mesurées suivantes ont un impact sur le **Outcome**, et les deux premières variables pourraient également avoir un impact sur **Exposure_stressor** (facteur de stress) : l'épaisseur de la paroi abdominale, la maladie artéritique, la cicatrisation des plaies, le protocole post-opératoire (traitement antibiotique) et la rééducation.

En ce qui concerne la direction du biais, les object properties signées n'ont pas été utilisées en raison de l'absence de contre-indication ou d'indication absolue. De plus, les propriétés *Increase* et *Decrease* n'ont pas été utilisées car les covariables causant le résultat augmentent également le résultat, et toutes les covariables causant l'exposition (à l'exception de l'épaisseur de la paroi abdominale) diminuent également l'exposition. Le chemin causal ne contient pas trop de nœuds, donc la direction du biais causée par un facteur de confusion est facile à anticiper sans inférences (Question n°5). Dans ce cas, le non ajustement serait une sous estimation de l'effet de l'exposition.

Le diagramme causal a été créé avec l'aide de l'interface de Protégé en renseignant le nom des instances et les relations qu'elles entretiennent entre elles.

4.2. Résultats

1) De la même manière que la première étape d'une recherche est la formulation de la question de recherche avec l'exposition d'intérêt et l'outcome, il faut spécifier dans OntoBioStat quelle instance est **Outcome** et laquelle est l'**Exposure_stressor**. Chacune de ces instances devra aussi être classées sous une variable théorique (respectivement **Condition** et **Intervention_Effect**).

2) Après cela, le raisonneur est activé une première fois ce qui entraîne la construction automatique d'un diagramme causal initial avec les causes nécessaires. Pour rappel, ces classes contiennent des instances génériques ce qui veut dire qu'elles peuvent être utilisées dans différents scénarios. En fonction de la classe théorique de l'**Outcome** et de l'**Exposure_stressor**, OntoBioStat utilise un sous ensemble de causes nécessaires pour créer le diagramme causal initial. Dans ce diagramme les variables nécessaires les instances génériques sont connectées soit à l'Outcome soit à l'**Exposure_stressor** avec des relations causales. Les variables nécessaires n'étaient pas utiles dans notre cas d'usage car, que ce soit l'**Exposure_stressor** ou l'**Outcome**, ce sont deux variables qui correspondent à des événements sporadiques (le traitement chirurgical n'est pas un traitement à long terme délivré par un pharmacien), et l'**Outcome** est cliniquement bruyant, il n'y a donc pas de biais de décalage diagnostic différentiel (lag bias) (c'est-à-dire qu'il n'y a pas de retard majeur entre l'apparition du trouble et son diagnostic). Il y aura donc peu ou pas de risque de biais de suivi ou de classification pour l'exposition ou le résultat. Cela signifie que les variables nécessaires ne seront pas en relation de causalité avec une autre variable que l'**Exposure_stressor** et l'**Outcome**. D'un point de vue formel, une variable qui influence le choix du traitement devrait causer la variable nécessaire **Indicated**, il serait donc légitime de vouloir utiliser les variables nécessaires.

3) Ensuite, les instances correspondantes aux **Covariates** sont spécifiées. Puis, leur contexte représenté par les **Meta_Variables** est spécifié. S'en suit la spécification des relations causales entre les différentes instances. Des informations additionnelles viennent enrichir la représentation via les data properties. Un total de 26 instances (15 **Covariates**, 11 **Metavariables** dont cinq **Missing_Value_Reasons**), 50 object properties et 37 data properties ont été spécifiées à l'aide de Protégé (Figure 67).

4) Le raisonneur est activé une deuxième fois, et permet de mettre en lumière les proxy et variables de confusion qui pourraient être intéressants à utiliser. Un total de 12 **Confounder**

ont été inférés. Des axiomes et des règles SWRL ont permis de répondre aux questions de compétence (voir Tableau 9). Les proxies suivants ont été détectés : Score ASA (pour la condition artéritique et l'instabilité hémodynamique initiale) ; Passage en réanimation (pour l'instabilité hémodynamique initiale) ; IMC (pour l'épaisseur de la paroi abdominale et la condition artéritique). Le Score ASA est proche de la condition artéritique (*isCauseof*), contrairement à l'IMC (*Share_ancestor*), c'est pourquoi le Score ASA a été préféré à l'IMC. En effet, nous supposons que plus une relation est éloignée dans la chaîne de la causalité plus elle a tendance à être faible. Cependant, comme le BMI est un bon proxy pour l'épaisseur abdominale, il a été utilisé en tant que proxy aussi. Le Score ASA avait peu de données manquantes, et dépendait du point de vue de l'auteur et de la collecte de données. Le Passage en réanimation avait des données manquantes en fonction du centre où les données étaient collectées (MAR : Données manquantes de manière aléatoire). L'IMC avait des données manquantes qui semblaient manquer complètement de manière aléatoire. Afin de comprendre les inférences liées au proxy, nous nous sommes appuyés sur la fonction d'inférence explicative. Un exemple d'explication est fourni dans la Figure 82, dans lequel l'instance Passage en réanimation est classée comme un **Proxy_Confounder** grâce à plusieurs assertions et règles SWRL (ligne 5).

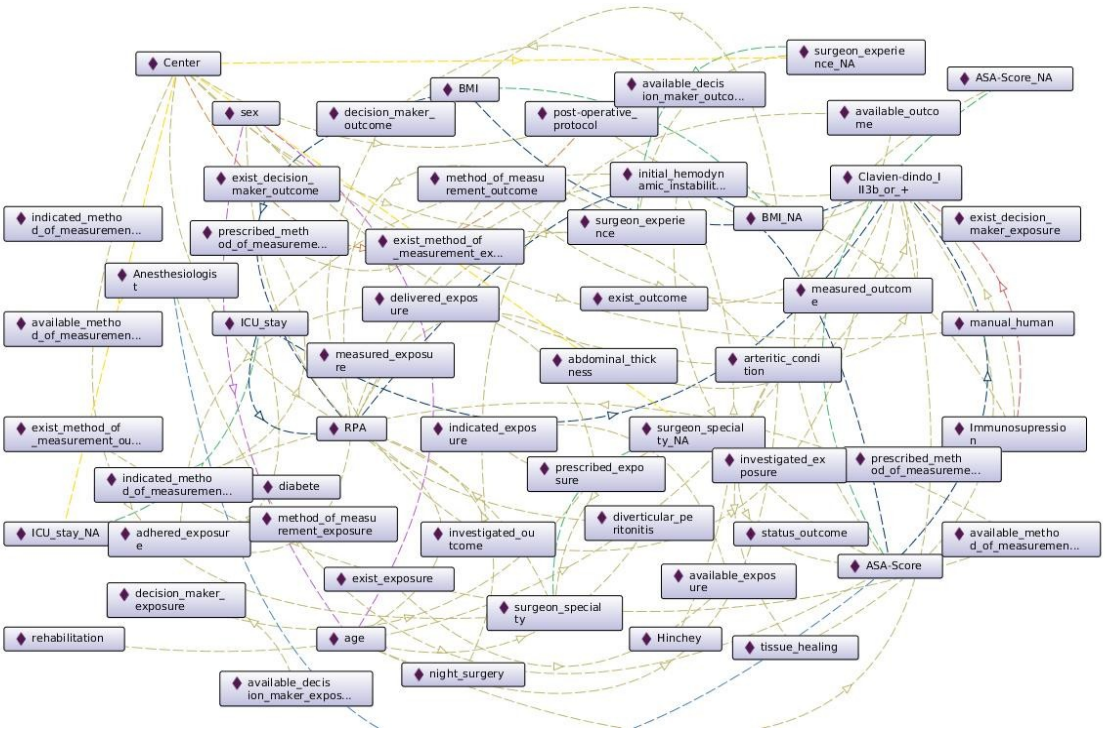


Figure 81. Graphe initial avec les instances et les object properties.

5-6) Le diagramme est modifié en substituant les variables par les proxies utilisés et le raisonneur est une dernière fois activé afin d'obtenir les inférences finales. Chacune d'entre elle est explicable grâce à la fonction explain inference. Analyse de sensibilité avec un ensemble alternatif de facteurs de confusion : Compte tenu des facteurs de confusion retrouvés ; l'Immunosuppression has**PossibleValue** Incomplete (variable de faible qualité) et la relation causale avec l'**Outcome** était en cours de validation. Les informations sur le chirurgien et l'emplacement du centre ont été considérées comme des facteurs de confusion en raison du lien causal hypothétique entre le protocole post-opératoire et l'**Outcome**. Ces variables sont considérées comme des facteurs de confusion en raison de ces liens causaux incertains, nous avons donc supprimé le lien causal hypothétique et en cours de validation et relancé le raisonneur afin d'obtenir un ensemble différent de facteurs de confusion.

Figure 82. Explain inference feature of the Protégé.

Tableau 9: Réponses aux questions de compétence

Questions de compétences	Réponses pour le cas d'usage
1) Faut-il exclure des patients qui ne seraient pas comparables aux autres ?	Non
2) Quelles sont les variables qui confondent le vrai effet causal entre exposition et critère	Age, sexe, maladie artéritique, ASA-Score, cicatrisation des plaies, l'instabilité hémodynamique, protocole post-opératoire,

Questions de compétences	Réponses pour le cas d'usage
de jugement ?	expérience et spécialité du chirurgien, épaisseur de la paroi abdominale, centre, immunosuppression.
3) Existe-t-il une interaction qui pourrait biaiser le vrai effet causal entre exposition et critère de jugement ?	Non concerné
4) Est-ce que le mécanisme responsable de la présence de données manquantes pourrait biaiser l'estimation du vrai effet causal entre exposition et critère de jugement ?	ASA-Score est MNAR, Passage en réanimation, expérience et spécialité du chirurgien sont MAR
5) Quelle est la direction des biais causés par des variables de confusion ?	Non concerné
6) Existe-t-ilil des variables proxies de variable de confusion ?	ASA-Score, Passage en réanimation, IMC
7) Quelle type de relation existe entre deux variables ?	1465 object properties ont été inférrées. Elles correspondent à Related_to et sa descendance : e.g: (i) Hinchey *isIndirectCauseof* RAP, (ii) ASA_score *Share_ancestor* with passage en réanimation, (iii) cicatrisation des plaies *Related_to* ASA-Score
8) Est-ce qu'il existe des biais qu'on ne peut pas corriger ?	Non

4.3. Discussion

Dans ce cas d'usage, OntoBioStat a fourni un canevas de construction riche, et a permis de mettre en évidence les biais, tout en expliquant au besoin les inférences fournies. Dans ce cas d'usage, toutes les fonctionnalités de OntoBioStat n'ont pas été utilisées telles que les variables nécessaires ou les object properties signées.

La saisie manuelle de toutes les classes, leurs liens et leurs data properties ralentit l'utilisation. De plus, les biostatisticiens ne sont pas habitués à l'utilisation d'un logiciel tel que Protégé. Les explications des inférences pourraient paraître opaques pour les débutants et longues à lire. OntoBioStat ne peut pas répondre de manière automatique à la question la plus importante qui est: 'Quel est le jeu minimal de variables qui permet d'obtenir le vrai effet causal?' avec seulement des règles SWRL ou des axiomes. Cela correspond au jeu minimal de variables à sélectionner qui devrait corriger tous les biais. En effet, l'analyse de chemin

causal, indispensable pour résoudre cette question en automatique, est seulement possible avec des fonctions JAVA non natives (si nous souhaitons rester dans l'environnement de Protégé). Il est tout à fait possible d'exporter l'ontologie sous R à partir d'un csv ou encore sous Neo4j un système de gestion de bases de données orientées graphes grâce à neosemantic.. Cependant, OntoBioStat permet bien une sélection manuelle, car il met en évidence les variables de confusion. Certaines des instances ne sont pas toujours impliquées dans un chemin causal entre exposition et critère de jugement. Ces nœuds non pertinents pourraient être supprimés grâce à une requête SPARQL ou une opération sur graphe. Cependant, les diagrammes causaux ontologiques resteraient difficiles à lire à l'oeil nu (sans logiciel) et supprimer de la connaissance déjà spécifiée devrait être considéré comme de la perte pure et simple (notamment en cas de réutilisation ou justification). À cause des limites évoquées plus haut, OntoBioStat devrait être considérée comme un outil pédagogique ou un filet de sécurité pour les utilisateurs peu compétents plutôt qu'un véritable système d'aide à la décision utilisable en routine.

Les outils tels que Dagitty pour la construction de DAG fournissent une interface très facile à prendre en main et rapide d'utilisation. L'utilisateur peut spécifier la structure du graphe en le 'dessinant' et obtient des renseignements basiques tels que les ancêtres communs de l'exposition et du critère de jugement, les ancêtres de l'exposition, ceux du critère de jugement. En outre, l'utilisateur visualise immédiatement si les variables sont mesurées ou non et si elles sont ajustées ou non. Pour finir, le jeu minimal de variables est facilement fourni. Cependant, Dagitty ne fournit pas de cadre formel de construction de diagramme causal contrairement à OntoBioStat, ni une richesse suffisante concernant les différents types de relations causales possibles (i.e., causes suffisantes, nécessaires, incertaines, signées, interaction, etc). Concernant la spécification des relations entre chaque variable, la construction par dessin à la main est rapide mais est soumise à l'erreur d'oubli alors que la construction par assertion oblige à se poser la question de relation causale entre chaque variable. Avec un nombre de variables important, un dessin ne permet pas de contrôler visuellement les erreurs potentielles alors que les inférences fournies par OntoBioStat permettent rapidement de détecter les relations causales aberrantes. Le monde ouvert de l'ontologie a des avantages et des inconvénients concernant *isCauseof :* « Monde ouvert » veut dire que si isCauseof n'est pas spécifié cela ne veut pas dire qu'il n'y a pas de relation causale entre les deux variables, alors que sur un graphe on considère que l'absence de flèche entre deux variables correspond à l'absence de relation causale. Dans OntoBioStat, il faut le spécifier (*notcauseof*) afin de pouvoir inférer les proxy potentiels. Dans Dagitty il est possible

de détecter les variables dites instrumentales mais pas les proxy. Dans ce cas d'usage, Dagitty n'aurait pas pu mettre en évidence les variables qui ont des données manquantes qui peuvent biaiser l'effet causal. De plus, toutes les informations supplémentaires contenues dans les data properties, telles que le pourcentage de données manquantes ou l'incomplétude d'une variable, ont une réelle plus-value par rapport à Dagitty et permettent d'apprécier la qualité des données pour faire un choix éclairé entre plusieurs jeux minimaux de variables.

Conclusion et Perspectives

La thèse présentée ici a exploré la création et l'utilisation d'OntoBioStat, une ontologie spécialement conçue pour répondre aux besoins de la sélection des variables en inférence causale pour la recherche biomédicale observationnelle. Les objectifs de cette thèse étaient de représenter toutes les informations nécessaires à la sélection des variables, de proposer un cadre formel pour aider à la construction de diagrammes causaux, d'analyser ces diagrammes causaux via des inférences, et enfin, d'assurer que tous les résultats inférés par l'ontologie soient explicables.

OntoBioStat joue un rôle essentiel en facilitant la représentation complète des informations cruciales pour une sélection éclairée des variables : (i) le contexte d'une étude avec les métavariables, (ii) les variables implicites avec les causes nécessaires et la classe théorique statut, (iii) l'incertitude de la connaissance elle-même avec des relations causales hypothétiques, (iv) les causes suffisantes avec les contre indications et les indications absolues, (v) les raisons de la présence de données manquante avec la relation *isCauseofNA* et la métavariable raison des données manquantes, (vi) la qualité des données avec des informations sur les données manquantes et les transformations préalables irréparables des données. OntoBioStat reste accessible grâce à son formalisme parcimonieux. Le nombre d'entités (classes, relations et dataproperties) a été limité au strict nécessaire pour éviter de l'alourdir inutilement et permet probablement d'améliorer son utilisabilité. Cette approche garantit que l'ontologie ne rebute pas les chercheurs utilisateurs de DAG, qui sont les principaux utilisateurs cibles.

OntoBioStat permet de construire un diagramme en fournissant un cadre de construction s'inspirant des composants basiques et avancés des DAG qui sont la référence pour une bonne sélection des variables. Il offre un cadre de construction de diagrammes causaux qui dépasse les limites du DAG traditionnel. Ce cadre est à la fois formel, explicite, systématique, et riche en contexte s'appuyant sur les informations cruciales citées plus haut.

Grâce à OntoBioStat, les chercheurs peuvent non seulement construire des diagrammes causaux, mais également effectuer des inférences automatiques sur ces diagrammes. L'analyse des diagrammes causaux au sein d'OntoBioStat est rigoureuse, grâce à l'utilisation d'un raisonneur basé sur des règles et des axiomes. Les règles SWRL et les axiomes permettent un raisonnement transparent et répondent à huit questions de compétences

essentielles pour guider la sélection des variables. Les enrichissements contextuels et explicites ouvrent la porte à de nouvelles inférences. Les différentes classes de variables de confusion, les proxy, les interactions, le mécanisme de génération des données manquantes, la causalité inverse, et les biais impossibles à corriger sont inférés grâce aux spécifications riches de l'utilisateur.

Dans le domaine de l'IA et de la causalité, et particulièrement en médecine l'explicabilité est devenue une préoccupation majeure, car les résultats automatisés peuvent souvent sembler ésotériques ou difficiles à interpréter. Grâce à l'interface de Protégé, il est facile d'obtenir une explication des inférences. Il est possible pour chaque inférence de retracer le raisonnement depuis les spécifications de l'utilisateur en passant par les règles ou axiomes utilisés, garantissant ainsi une compréhension claire de la manière dont les conclusions ont été tirées. Cette transparence est essentielle pour les chercheurs, car elle renforce la confiance dans les résultats automatisés. OntoBioStat se comporte comme un assistant méthodologique complet pour la sélection des variables dans l'inférence causale, prenant en charge un large éventail de missions, de la représentation fine de la question de recherche avec les classes théoriques à l'interprétation des résultats grâce à son explicabilité.

En conclusion, cette thèse a atteint ses objectifs en développant OntoBioStat, une ontologie pour la sélection des variables en inférence causale. OntoBioStat offre une représentation holistique de la connaissance d'une étude, un cadre formel pour la construction de diagrammes causaux, une analyse automatique des diagrammes causaux, une explicabilité des résultats, et peut être appliquée à des problématiques réelles. C'est un outil original pour les chercheurs en épidémiologie et en statistiques.

Perspectives

Automatisation de la saisie de données : Une direction de recherche importante consiste à explorer des méthodes pour automatiser davantage la saisie de données dans OntoBioStat. Actuellement, la création d'instances d'ontologie nécessite souvent une saisie manuelle. Le développement de techniques d'extraction automatique d'informations à partir de bases de données ou de textes scientifiques pourrait accélérer ce processus. Il serait aussi envisageable de réutiliser des graphes de connaissance existants sous la forme de DAG ou d'ontologies pour pré-remplir des instances et leurs relations causales comme Rodríguez-García, M et al., 2018 et Malec SA et al., 2023.

Intégration avec des outils d'analyse de données et de construction: OntoBioStat pourrait être intégrée à des outils d'analyse de données existants, tels que le logiciel statistique R ou

Dagitty pour rendre l'ontologie plus accessible aux chercheurs non spécialisés en ontologie. Cela permettrait aux chercheurs d'appliquer directement les connaissances causales de l'ontologie dans leurs analyses. Elle pourrait inclure des fonctionnalités de recherche avancée, des outils de visualisation de diagrammes causaux, et des guides interactifs pour la sélection des variables. Par ailleurs, il existe un besoin pressant d'améliorer l'utilisabilité d'OntoBioStat et de fournir un ensemble minimal de variables essentielles pour corriger les biais. Compte tenu de la prévalence de R parmi les biostatisticiens, l'utilisation conjointe de Dagitty et d'OntoBioStat pourrait s'avérer particulièrement bénéfique. Une fois le graphe construit grâce à OntoBioStat, il pourrait être analysé à l'aide du package Dagitty. Cette approche permettrait d'identifier différents ensembles minimaux de variables pertinents, le tout accompagné d'explications fournies par OntoBioStat. Grâce à la représentation des liens causaux incertains, de la gestion des données manquantes, de la qualité des données et de leur source, cette méthode aiderait à prendre des décisions éclairées dans le choix des variables à inclure dans l'analyse, améliorant ainsi la robustesse des résultats d’une étude donnée. Par ailleurs, Fluenteditor est un éditeur interopérable avec Protégé et a produit un package rOntorion qui permet d’utiliser l’ontologie dans R. Hanly M et al., 2023 ont créé daggle app qui correspond à une application en R shiny s’appuyant entre autre sur Dagitty.

Extension vers d'autres domaines : Bien qu'OntoBioStat ait été conçue pour l’inférence causale dans le cadre de la recherche biomédicale, son potentiel peut s'étendre à d'autres domaines de la recherche scientifique non médicale. Des adaptations de l'ontologie pourraient être envisagées pour des domaines tels que la biologie, la sociologie ou l’économie.

Mesurer l’impact de son utilisation : Des études empiriques visant à valider l'efficacité d'OntoBioStat dans la sélection des variables et l'analyse des résultats pourraient être entreprises. Cela impliquerait de comparer les résultats obtenus en utilisant OntoBioStat avec ceux d'autres approches ou outils existants.

Élargissement de la communauté d'utilisateurs : Encourager la communauté de recherche à adopter OntoBioStat en tant qu'outil standard pour la sélection des variables pourrait accélérer son développement et sa pertinence. Des ateliers de formation, des tutoriels et des collaborations interdisciplinaires pourraient favoriser son adoption. Par la même occasion, son impact chez les apprenants pourrait être évalué.

Bibliographie

1. Ankan A, Wortel IMN, Textor J. Testing Graphical Causal Models Using the R Package “dagitty”. Current Protocols. févr 2021;1(2). Disponible sur: https://onlinelibrary.wiley.com/doi/10.1002/cpz1.45

2. Armitage P. Fisher, Bradford Hill, and randomization. International Journal of Epidemiology. déc 2003;32(6):925-8.

3. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. Xue Y, éditeur. PLoS ONE. 29 avr 2016;11(4):e0154556.

4. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci USA. 5 juill 2016;113(27):7345-52.

5. Barnard-Mayers R, Childs E, Corlin L, Caniglia EC, Fox MP, Donnelly JP, et al. Assessing knowledge, attitudes, and practices towards causal directed acyclic graphs: a qualitative research project. Eur J Epidemiol. juill 2021;36(7):659-67.

6. Bärnighausen T, Røttingen JA, Rockers P, Shemilt I, Tugwell P. Quasi-experimental study designs series—paper 1: introduction: two historical lineages. Journal of Clinical Epidemiology. sept 2017;89:4-11.

7. Behnaz A, Bandara M, Rabhi FA, Peat M. A Statistical Learning Ontology for Managing Analytics Knowledge. In: Mehandjiev N, Saadouni B, éditeurs. Enterprise Applications, Markets and Services in the Finance Industry [Internet]. Cham: Springer International Publishing; 2019 [cité 10 oct 2023]. p. 180-94. (Lecture Notes in Business Information Processing; vol. 345). Disponible sur: http://link.springer.com/10.1007/978-3-030-19037-8_12

8. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 1 sept 2017;2(1):6-10.

9. Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data.*,†. International Journal of Epidemiology. avr 2014;43(2):511-5.

10.

Besnard Ph, Cordier MO, Moinard Y. Ontology-based inference for causal explanation1. ICA. 31 juill 2008;15(4):351-67.

11. Bhandari PM, Levis B, Neupane D, Patten SB, Shrier I, Thombs BD, et al. Data-driven methods distort optimal cutoffs and accuracy estimates of depression screening tools: a simulation study using individual participant data. Journal of Clinical Epidemiology. sept 2021;137:137-47.

12. Bock A, Fokoue P, Haase R. OWL 2 Web Ontology Language, W3C recommendation.

13. Borst P, Akkermans H, Top J. Engineering ontologies. International Journal of Human-Computer Studies. févr 1997;46(2-3):365-406.

14. Breitling LP, Duan C, Dragomir AD, Luta G. Using dagR to identify minimal sufficient adjustment sets and to simulate data based on directed acyclic graphs. International Journal of Epidemiology. 6 janv 2022;50(6):1772-7.

15. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med. déc 2008;3(1):17.

16. Cadarette SM, Maclure M, Delaney JAC, Whitaker HJ, Hayes KN, Wang SV, et al. Control yourself: ISPE-ENDORSED guidance in the application of SELF-CONTROLLED study designs in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. juin 2021;30(6):671-84.

17. Ceusters W. An information artifact ontology perspective on data collections and associated representational artifacts. Stud Health Technol Inform. 2012;180:68-72.

18. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 9 janv 2013;346(jan08 15):e7586-e7586.

19. Chandrasekaran B, Josephson JR, Benjamins VR. What are ontologies, and why do we need them? IEEE Intell Syst. janv 1999;14(1):20-6.

20. Chari S, Qi M, Agu NN, Seneviratne O, McCusker JP, Bennett KP, et al. Making Study Populations Visible Through Knowledge Graphs. In: Ghidini C, Hartig O, Maleshkova M, Svátek V, Cruz I, Hogan A, et al., éditeurs. The Semantic Web – ISWC 2019 [Internet]. Cham: Springer International Publishing; 2019 [cité 29 mars 2023]. p. 53-68. (Lecture Notes in

Computer Science; vol. 11779). Disponible sur: https://link.springer.com/10.1007/978-3-030-30796-7_4

21. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of *p* -values. R Soc open sci. nov 2014;1(3):140216.

22. Cowell LG, Smith B. Infectious Disease Ontology. In: Sintchenko V, éditeur. Infectious Disease Informatics [Internet]. New York, NY: Springer New York; 2010 [cité 28 mars 2023]. p. 373-95. Disponible sur: http://link.springer.com/10.1007/978-1-4419-1327-2_19

23. de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, et al. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. juin 2009;42(3):530-9.

24. Desboulets L. A Review on Variable Selection in Regression Analysis. Econometrics. 23 nov 2018;6(4):45.

25. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. Journal of Clinical Epidemiology. févr 2022;142:264-7.

26. Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented Backward Elimination: A Pragmatic and Purposeful Way to Develop Statistical Models. Olivier J, éditeur. PLoS ONE. 21 nov 2014;9(11):e113677.

27. Elm EV, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 20 oct 2007;335(7624):806-8.

28. Falbo R de A. The Systematic Approach for Building Ontologies. 2014;

29. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). nov 2008;70(5):849-911.

30. Ferguson KD, McCann M, Katikireddi SV, Thomson H, Green MJ, Smith DJ, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. International Journal of Epidemiology. 1 févr 2020;49(1):322-9.

31.

Fernández-López M, Gómez-Pérez A, Juristo N. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In: Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series. Stanford University; 1997.

32. Flom PL, Development N, Institutes R, York N. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. 2007;7.

33. for TG2 of the STRATOS initiative, Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagn Progn Res. déc 2020;4(1):3, s41512-020-00074-3.

34. Franklin JDS, Chari S, Foreman MA, Seneviratne O, Gruen DM, McCusker JP, et al. Knowledge Extraction of Cohort Characteristics in Research Publications. AMIA Annu Symp Proc. 2020;2020:462-71.

35. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology. 1 avr 2011;173(7):761-7.

36. Galton A. States, processes and events, and the ontology of causal relations. Frontiers in Artificial Intelligence and Applications. janv 2012;239:279-92.

37. Glimm B, Horrocks I, Motik B, Stoilos G, Wang Z. HermiT: An OWL 2 Reasoner. J Autom Reasoning. oct 2014;53(3):245-69.

38. Glymour MM, Weuve J, Chen JT. Methodological Challenges in Causal Research on Racial and Ethnic Patterns of Cognitive Trajectories: Measurement, Selection, and Bias. Neuropsychol Rev. sept 2008;18(3):194-213.

39. Gómez-Pérez A. Ontology Evaluation. In: Staab S, Studer R, éditeurs. Handbook on Ontologies [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004 [cité 28 mars 2023]. p. 251-73. Disponible sur: http://link.springer.com/10.1007/978-3-540-24750-0_13

40. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. Seminars in Hematology. juill 2008;45(3):135-40.

41. Greenland S. Invited Commentary: Variable Selection versus Shrinkage in the Control of Multiple Confounders. American Journal of Epidemiology. 12 déc 2007;167(5):523-9.

42.
Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. janv 1999;10(1):37-48.
43.
Greenland S, Robins JM. Identifiability, Exchangeability, and Epidemiological Confounding. Int J Epidemiol. 1986;15(3):413-9.
44.
Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet. 19 janv 2002;359(9302):248-52.
45.
Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition. juin 1993;5(2):199-220.
46.
Gruber TR. Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies. nov 1995;43(5-6):907-28.
47.
Gruninger M, Fox M. Methodology for the Design and Evaluation of Ontologies. International Joint Conference on Artificial Intelligence [Internet]. 1995; Disponible sur: https://www.semanticscholar.org/paper/Methodology-for-the-Design-and-Evaluation-of-Gruninger/497abc0ddace6a7772a5f5a3edb3d7b751476755
48.
Guarino N, éditeur. Formal ontology in information systems: proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy. Amsterdam ; Washington, DC : Tokyo: IOS Press ; Omsha; 1998. 337 p. (Frontiers in artificial intelligence and applications).
49.
Gündel M, Younesi E, Malhotra A, Wang J, Li H, Zhang B, et al. HuPSON: the human physiology simulation ontology. J Biomed Sem. 2013;4(1):35.
50.
G�mez-P�rez A. Evaluation of ontologies. Int J Intell Syst. mars 2001;16(3):391-409.
51.
Hamaker HC. On multiple regression analysis. Statistica Neerland. mars 1962;16(1):31-56.
52.
Hanly M, Brew BK, Austin A, Jorm L. Software Application Profile: The daggle app—a tool to support learning and teaching the graphical rules of selecting adjustment variables using directed acyclic graphs. International Journal of Epidemiology. 5 oct 2023;52(5):1659-64.
53.
Hernán MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. Am J Public Health. mai 2018;108(5):616-9.

54.
Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias: Epidemiology. sept 2004;15(5):615-25.
55.
Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. Am J Epidemiol. 15 avr 2016;183(8):758-64.
56.
Hernán M, Robins JM. Causal inference. Boca Raton: Chapman & Hall/CRC; 2021. 352 p. (Chapman & Hall/CRC monographs on statistics & applied probability).
57.
Hoehndorf R, Gkoutos GV, Schofield PN. Datamining with Ontologies. In: Carugo O, Eisenhaber F, éditeurs. Data Mining Techniques for the Life Sciences [Internet]. New York, NY: Springer New York; 2016 [cité 28 mars 2023]. p. 385-97. (Methods in Molecular Biology; vol. 1415). Disponible sur: http://link.springer.com/10.1007/978-1-4939-3572-7_19
58.
Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 7 mars 2014;348(mar07 3):g1687-g1687.
59.
Holmberg MJ, Andersen LW. Adjustment for Baseline Characteristics in Randomized Clinical Trials. JAMA. 6 déc 2022;328(21):2155.
60.
Horrocks I. SWRL: A Semantic Web Rule Language Combining OWL and RuleML [Internet]. 2004. Disponible sur: http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/
61.
Ioannidis JPA. Exposure-wide epidemiology: revisiting Bradford Hill. Stat Med. 20 mai 2016;35(11):1749-62.
62.
Julious SA, Mullee MA. Confounding and Simpson’s paradox. BMJ. 3 déc 1994;309(6967):1480-1.
63.
Kahn CE. Ontology-based diagnostic decision support in radiology. Stud Health Technol Inform. 2014;205:78-82.
64.
Kahn CE. Transitive closure of subsumption and causal relations in a large ontology of radiological diagnosis. Journal of Biomedical Informatics. juin 2016;61:27-33.
65.

Kalyanpur A, Parsia B, Sirin E, Grau BC, Hendler J. Swoop: A Web Ontology Editing Browser. Journal of Web Semantics. juin 2006;4(2):144-53.

66. Kestenbaum B. Population, Exposure, and Outcome. In: Epidemiology and biostatistics: practice problem workbook. New York, NY: Springer Berlin Heidelberg; 2018.

67. Koch B, Vock DM, Wolfson J. Covariate selection with group lasso and doubly robust estimation of causal effects: GLiDeR. Biom. mars 2018;74(1):8-17.

68. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ. 14 nov 2018;k3532.

69. Lawrance R, Degtyarev E, Griffiths P, Trask P, Lau H, D'Alessio D, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? J Patient Rep Outcomes. déc 2020;4(1):68.

70. Lee J, Kamdar BB, Bergstrom J, Murphy TE, Gill TM. Modeling success: How to work effectively with your biostatistician. J American Geriatrics Society. août 2022;70(8):2449-54.

71. Lee S, Zhu Y. Confounder Selection via Support Intersection [Internet]. arXiv; 2019 [cité 28 mars 2023]. Disponible sur: http://arxiv.org/abs/1912.11652

72. Lénart M. SKOS, un langage de représentation de schémas de concepts. Documentaliste-Sciences de l'Information. 2007;44(1):75.

73. Lewis D. Causation. The Journal of Philosophy. 11 oct 1973;70(17):556.

74. Li F, Du J, He Y, Song HY, Madkour M, Rao G, et al. Time event ontology (TEO): to support semantic representation and reasoning of complex temporal relations of clinical events. Journal of the American Medical Informatics Association. 1 juill 2020;27(7):1046-56.

75. Liao H, Lynn HS. A survey of variable selection methods in two Chinese epidemiology journals. BMC Med Res Methodol. déc 2010;10(1):87.

76. Lopez PM, Subramanian SV, Schooling CM. Effect measure modification conceptualized using selection diagrams as mediation by mechanisms of varying population-level relevance. Journal of Clinical Epidemiology. sept 2019;113:123-8.

77.
Malec SA, Taneja SB, Albert SM, Elizabeth Shaaban C, Karim HT, Levine AS, et al. Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: A use case studying depression as a risk factor for Alzheimer's disease. Journal of Biomedical Informatics. juin 2023;142:104368.
78.
Mansournia MA, Nazemipour M, Etminan M. Causal diagrams for immortal time bias. International Journal of Epidemiology. 10 nov 2021;50(5):1405-9.
79.
Matentzoglu N, Malone J, Mungall C, Stevens R. MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semant. déc 2018;9(1):6.
80.
Miettinen OS, Cook EF. CONFOUNDING: ESSENCE AND DETECTION1. American Journal of Epidemiology. oct 1981;114(4):593-603.
81.
Mohan K, Pearl J. Graphical Models for Processing Missing Data. Journal of the American Statistical Association. 3 avr 2021;116(534):1023-37.
82.
Musen MA. The protégé project: a look back and a look forward. AI Matters. 16 juin 2015;1(4):4-12.
83.
Nayak A, Božić B, Longo L. An Ontological Approach for Recommending a Feature Selection Algorithm. In: Di Noia T, Ko IY, Schedl M, Ardito C, éditeurs. Web Engineering [Internet]. Cham: Springer International Publishing; 2022 [cité 28 mars 2023]. p. 300-14. (Lecture Notes in Computer Science; vol. 13362). Disponible sur: https://link.springer.com/10.1007/978-3-031-09917-5_20
84.
Nilsson A, Bonander C, Strömberg U, Björk J. A directed acyclic graph for interactions. International Journal of Epidemiology. 17 mai 2021;50(2):613-9.
85.
Noy N, McGuiness D. Ontology Development 101: A Guide to Creating Your First Ontology. 2001.
86.
O'Connor M, Tu S, Nyulas C, Das A, Musen M. Querying the Semantic Web with SWRL. In: Paschke A, Biletskiy Y, éditeurs. Advances in Rule Interchange and Applications [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007 [cité 28 mars 2023]. p. 155-9. (Lecture Notes in Computer Science; vol. 4824). Disponible sur: http://link.springer.com/10.1007/978-3-540-75975-1_13

87.
Okhmatovskaia A, Shaban-Nejad A, Lavigne M, Buckeridge DL. Addressing the challenge of encoding causal epidemiological knowledge in formal ontologies: a practical perspective. Stud Health Technol Inform. 2014;205:1125-9.
88.
Panov P, Soldatova L, Džeroski S. Ontology of core data mining entities. Data Min Knowl Disc. sept 2014;28(5-6):1222-65.
89.
Panov P, Soldatova LN, Džeroski S. Generic ontology of datatypes. Information Sciences. févr 2016;329:900-20.
90.
Pearl J. [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention. Statist Sci [Internet]. 1 août 1993 [cité 28 mars 2023];8(3). Disponible sur: https://projecteuclid.org/journals/statistical-science/volume-8/issue-3/Bayesian-Analysis-in-Expert-Systems--Comment--Graphical-Models/10.1214/ss/1177010894.full
91.
Pearl J. Causal Inference in the Health Sciences: A Conceptual Introduction. Health Services and Outcomes Research Methodology. 2001;2(3/4):189-220.
92.
Pearl J, Mackenzie D. The book of why: the new science of cause and effect. London: Penguin Books; 2019. 418 p. (Penguin science).
93.
Pesquita C, Ferreira JD, Couto FM, Silva MJ. The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. J Biomed Sem. 2014;5(1):4.
94.
Porzel R, Malaka R. A Task-based Approach for Ontology Evaluation. ECAI Workshop on Ontology Learning and Population. 2004;
95.
Poveda-Villalón M, Gómez-Pérez A, Suárez-Figueroa MC. OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. International Journal on Semantic Web and Information Systems. 1 avr 2014;10(2):7-34.
96.
Poveda-Villalón M, Suárez-Figueroa MC, Gómez-Pérez A. Validating Ontologies with OOPS! In: ten Teije A, Völker J, Handschuh S, Stuckenschmidt H, d'Acquin M, Nikolov A, et al., éditeurs. Knowledge Engineering and Knowledge Management [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 [cité 28 mars 2023]. p. 267-81. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al. Lecture Notes in Computer Science; vol. 7603). Disponible sur: http://link.springer.com/10.1007/978-3-642-33876-2_24

97.
Pressat-Laffouilhère T, Jouffroy R, Leguillou A, Kerdelhue G, Benichou J, Gillibert A. Variable selection methods were poorly reported but rarely misused in major medical journals: Literature review. Journal of Clinical Epidemiology. nov 2021;139:12-9.
98.
Ramspek CL, Steyerberg EW, Riley RD, Rosendaal FR, Dekkers OM, Dekker FW, et al. Prediction or causality? A scoping review of their conflation within current observational research. Eur J Epidemiol. sept 2021;36(9):889-98.
99.
Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):A12-13.
100.
Robins JM, Morgenstern H. The foundations of confounding in epidemiology. Computers & Mathematics with Applications. 1987;14(9-12):869-916.
101.
Rodríguez-García MÁ, Hoehndorf R. Inferring ontology graph structures using OWL reasoning. BMC Bioinformatics. déc 2018;19(1):7.
102.
Rothman KJ, Greenland S. Causation and Causal Inference in Epidemiology. Am J Public Health. juill 2005;95(S1):S144-50.
103.
Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.
104.
Rutter M. Epidemiological methods to tackle causal questions. International Journal of Epidemiology. 1 févr 2009;38(1):3-6.
105.
Sackett DL. Bias in analytic research. Journal of Chronic Diseases. janv 1979;32(1-2):51-63.
106.
Schisterman EF, Cole SR, Platt RW. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. Epidemiology. juill 2009;20(4):488-95.
107.
Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. Epidemiology. juill 2009;20(4):512-22.
108.
Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 23 mars 2010;340(mar23 1):c332-c332.

109.
Shimonovich M, Pearce A, Thomson H, Keyes K, Katikireddi SV. Assessing causality in epidemiology: revisiting Bradford Hill to incorporate developments in causal thinking. Eur J Epidemiol. sept 2021;36(9):873-87.
110.
Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of Clinical Research (OCRe): An informatics foundation for the science of clinical research. Journal of Biomedical Informatics. déc 2014;52:78-91.
111.
Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y. Pellet: A practical OWL-DL reasoner. Journal of Web Semantics. juin 2007;5(2):51-3.
112.
Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46.
113.
Souza ÉF de, Falbo R de A, Vijaykumar NL. ROoST: Reference Ontology on Software Testing. AO. 13 mars 2017;12(1):59-90.
114.
Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 28 août 2019;l4898.
115.
Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 12 oct 2016;i4919.
116.
Stovitz SD, Shrier I. Causal inference for clinicians. BMJ EBM. juin 2019;24(3):109-12.
117.
Suissa S. Immortal Time Bias in Pharmacoepidemiology. American Journal of Epidemiology. 7 janv 2008;167(4):492-9.
118.
Sure Y, Angele J, Staab S. OntoEdit: Multifaceted Inferencing for Ontology Engineering. In: Spaccapietra S, March S, Aberer K, éditeurs. Journal on Data Semantics I [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003 [cité 28 mars 2023]. p. 128-52. (Goos G, Hartmanis J, van Leeuwen J. Lecture Notes in Computer Science; vol. 2800). Disponible sur: http://link.springer.com/10.1007/978-3-540-39733-5_6
119.
Sure Y, Erdmann M, Angele J, Staab S, Studer R, Wenke D. OntoEdit: Collaborative Ontology Development for the Semantic Web. In: Horrocks I, Hendler J, éditeurs. The

Semantic Web — ISWC 2002 [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2002 [cité 28 mars 2023]. p. 221-35. (Goos G, Hartmanis J, van Leeuwen J. Lecture Notes in Computer Science; vol. 2342). Disponible sur: http://link.springer.com/10.1007/3-540-48005-6_18

120. Suzuki E, Shinozaki T, Yamamoto E. Causal Diagrams: Pitfalls and Tips. Journal of Epidemiology. 5 avr 2020;30(4):153-62.

121. Suzuki E, Tsuda T, Mitsuhashi T, Mansournia MA, Yamamoto E. Errors in causal inference: an organizational schema for systematic error and random error. Annals of Epidemiology. nov 2016;26(11):788-793.e1.

122. Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. Eur J Epidemiol. août 2019;34(8):725-30.

123. Tambassi T. Completeness in Information Systems Ontologies. Axiomathes. déc 2022;32(S2):215-24.

124. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. International Journal of Epidemiology. 17 mai 2021;50(2):620-32.

125. 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. 15 janv 2017;dyw341.

126. Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet. oct 2019;51(10):1429-33.

127. Thompson B. Why Won’t Stepwise Methods Die? Measurement and Evaluation in Counseling and Development. janv 1989;21(4):146-8.

128. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). janv 1996;58(1):267-88.

129.

Tsarkov D, Horrocks I. FaCT++ Description Logic Reasoner: System Description. In: Furbach U, Shankar N, éditeurs. Automated Reasoning [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006 [cité 28 mars 2023]. p. 292-7. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al. Lecture Notes in Computer Science; vol. 4130). Disponible sur: http://link.springer.com/10.1007/11814771_26

130. Uschold M, Gruninger M. Ontologies: principles, methods and applications. The Knowledge Engineering Review. juin 1996;11(2):93-136.

131. VanderWeele TJ, Hernan MA. Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs. American Journal of Epidemiology. 15 juin 2012;175(12):1303-10.

132. Vanderweele TJ, Staudt N. Causal diagrams for empirical legal research: a methodology for identifying causation, avoiding bias and interpreting results. Law, Probability and Risk. 1 déc 2011;10(4):329-54.

133. VanderWeele TJ. Mediation and mechanism. Eur J Epidemiol. mai 2009;24(5):217-24.

134. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol. 15 mars 2019;34(3):211-9.

135. VanderWeele TJ, Hernán MA, Robins JM. Causal Directed Acyclic Graphs and the Direction of Unmeasured Confounding Bias. Epidemiology. sept 2008;19(5):720-8.

136. VanderWeele TJ, Robins JM. Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. Epidemiology. sept 2007;18(5):561-8.

137. VanderWeele TJ, Robins JM. Signed Directed Acyclic Graphs for Causal Inference. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1 janv 2010;72(1):111-27.

138. VanderWeele TJ, Shpitser I. A New Criterion for Confounder Selection. Biometrics. déc 2011;67(4):1406-13.

139. Viet SM, Falman JC, Merrill LS, Faustman EM, Savitz DA, Mervish N, et al. Human Health Exposure Analysis Resource (HHEAR): A model for incorporating the exposome into health studies. International Journal of Hygiene and Environmental Health. juin 2021;235:113768.

140.

Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. déc 2009;24(12):733-6.

141.

Weinberg CR. Can DAGs Clarify Effect Modification? Epidemiology. sept 2007;18(5):569-72.

142.

Westreich D, Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. American Journal of Epidemiology. 15 févr 2013;177(4):292-8.

143.

Westreich D. Berkson's Bias, Selection Bias, and Missing Data. Epidemiology. janv 2012;23(1):159-64.

144.

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour. Journal of Animal Ecology. sept 2006;75(5):1182-9.

145.

Witte J, Didelez V. Covariate selection strategies for causal inference: Classification and comparison. Biom J. sept 2019;61(5):1270-89.

146.

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J Royal Statistical Soc B. févr 2006;68(1):49-67.

147.

Zapf A, Huebner M, Rauch G, Kieser M. What makes a biostatistician? Statistics in Medicine. 20 févr 2019;38(4):695-701.

148.

Zheng J, Harris MR, Masci AM, Lin Y, Hero A, Smith B, et al. The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis. J Biomed Semant. déc 2016;7(1):53.

149.

Zheng Z, Baifan Z. Towards A Statistic Ontology for Data Analysis in Smart Manufacturing. International Workshop on the Semantic Web. 2022;

150.

Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B. avr 2005;67(2):301-20.

151.

Evaluation and assessment of knowledge sharing technology.

Annexes

Liste des communications et publications publiées ou soumis en lien avec la thèse de science

Pressat-Laffouilhère T, Jouffroy R, Leguillou A, Kerdelhue G, Benichou J, Gillibert A. Variable selection methods were poorly reported but rarely misused in major medical journals: Literature review. J Clin Epidemiol. 2021 Jul 16;139:12-19.

Pressat Laffouilhère T, Grosjean J, Bénichou J, Darmoni SJ, Soualmia LF (2021). Ontological Models Supporting Covariates Selection in Observational Studies. Stud Health Technol Inform. 2021 May 27;281:1095-1096. **MIE, 2021**

Pressat Laffouilhère T, Grosjean J, Bénichou J, Darmoni SJ, Soualmia LF (2022). OntoBioStat: Supporting Causal Diagram Design and Analysis. Stud Health Technol Inform. 2022 May 25;294:302-306. **MIE 2022**

Pressat Laffouilhère, T, Grosjean J, Pinson J, Darmoni SJ, Leveque E, Lanoy E, Bénichou J, Soualmia LF.(2022). Ontological Representation of Causal Relations for a Deep Understanding of Associations Between Variables in Epidemiology. In: Michalowski, M., Abidi, S.S.R., Abidi, S. (eds) Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science(), vol 13263. Springer, Cham. https://doi.org/10.1007/978-3-031-09342-5_5 **AIME 2022**

Pressat Laffouilhère, T, Grosjean J, Pinson J, Darmoni SJ, Leveque E, Lanoy E, Bénichou J, Soualmia LF. OntoBioStat: an ontology for covariate selection within the framework of causal inference in observational studies. Soumis à la revue Computers in biology and medicine.

Publications

Variable selection methods were poorly reported but rarely misused in major medical journals : Literature review

T Pressat-Laffouilhère[1,2,3*], R Jouffroy[4], A Leguillou[5], G Kerdelhue[2], J Benichou[1,6], A Gillibert[1].

1- CHU Rouen, Department of Biostatistics, F-76000 Rouen, France
2- CHU Rouen, Department of Biomedical Informatics, F-76000 Rouen, France
3- Normandie Univ, UNIROUEN, LITIS EA 4108, F-76000 Rouen, France
4- Intensive care unit, Anaesthesiology Department and SAMU of Paris, Necker University Hospital, Assistance Publique Hôpitaux de Paris and Paris Descartes university, 75015 Paris, France
5- Reims University Hospital, Reims, France
6- Inserm U 1018, University of Rouen and University Paris-Saclay, France
*:t.pressat@chu-rouen.fr; Rouen University Hospital 37 Boulevard Gambetta, 76000 Rouen, France

Abstract

Objective: This work presents a review of the literature on reporting, practice and misuse of *knowledge-based* and data-driven variable selection methods, in five highly cited medical journals, considering recoding and interaction unlike previous reviews.

Study Design and Setting: Original observational studies with a predictive or explicative research question with multivariable analyses published in NEJM, Lancet, JAMA, BMJ and AIM between 2017 and 2019 were searched. Article screening was performed by a single reader, data extraction was performed by two readers and a third reader participated in case of disagreement. The use of data-driven variable selection methods in *causal explicative* questions was considered as misuse.

Results: 488 articles were included. The variable selection method was unclear in 234 (48%) articles, data-driven in 78 (16%) articles and *knowledge-based* in 176 (36%) articles. The most common *data-driven* methods were: Univariate selection (n=22, 4.5%) and model comparisons or testing for interaction (n=17, 3.5%). Data-driven methods were misused in 51 (10.5%) of articles.

Conclusion: Overall reporting of variable selection methods is insufficient. *Data-driven* methods seem to be used only in a minority of articles of the big five medical journals.

Keywords: Review, Observational study, covariate selection, variable selection, reporting, data-driven
Running title: Variable selection methods: reporting, practice, misuse.

Article Wordcount = 2937

Introduction

Variable selection in multivariate analysis is a wide-ranging and still controversial issue affecting all quantitative areas of biomedical research (e.g. medecine, epidemiology, social sciences, etc.). Relevant variable selection is critical in observational studies to address confounding issues in explanatory models and reach high predictive performance in predictive models.

There are two general approaches to variable selection, (i) knowledge-based approaches based on published knowledge or expert opinion that may be synthesized in causal diagrams [1], and (ii) data-driven methods. There is a wide range of data-driven methods, including backward elimination, forward selection [2] and stepwise, Least Absolute Shrinkage and Selection Operator (LASSO) [3], Elasticnet [4], the “change in estimate”, augmented backward elimination [5], which combines backward elimination and “change in estimate” (more details in supplementary). These methods have been recently reviewed by Desboulet and al [6], Witte J and al [7]and Heinze G and al [8].

Data-driven methods are generally appropriate for building *predictive models* of a prognostic or diagnostic nature but are more questionable for *explanatory models* in etiologic research (*e.g.* models pertaining to the assessment of causal risk factors or treatment effects). Indeed, the latter requires making inference on the exposure (or treatment) coefficient of the model, with statistical tests or confidence intervals that are affected by bias when data-driven methods are used. Indeed, usual Wald, Rao’s score and likelihood ratio have been shown to be biased with too narrow confidence intervals and too high estimates in case of data-driven pre-selection of variables [9]. Moreover, data-driven methods may inappropriately adjust on mediation variables or omit relevant confounders, biasing inference on the main causal effect in explanatory models. In case of predictive models these adjustments are less deleterious

because the aim is to obtain high predictive performance whatever variable is used in the model.

Although STROBE guidelines [10] encourage authors to clearly define which variables are potential confounders and to thoroughly describe statistical methods in their papers, this is not always done [11] and many statistical choices are left unexplained, sometimes because of insufficient space.

Walter *and Timeier* reported frequencies of variable selection methods in four major epidemiological journals in 2008 and found widespread use of stepwise selection methods and frequent insufficient or missing reporting of methods used [12]. Talbot *et al* updated the review in 2015 and found a lower use of stepwise [13]. Of note, these reviews did not report on practices regarding assessment and inclusion of interaction terms, recoding of variables (e.g. the variable 'age' may be included as a continuous, categorical [18-25][26-35][36-45], or polynomial variable with $age+age^2+age^3$), or variable selection in sensitivity analyses.

Practices have evolved and may be different in medical journals with the highest impact factor, considered more influential as they are often seen as the cream of the crop. The goal of this study was therefore to review articles published from 2017 to 2019 in the five medical journals with the highest impact factor in the "Medicine, general & internal" category according to the journal citation reports of 2016 [14], often referred to as the big-five medical journals (i.e., New England Journal of Medicine, The Lancet, Journal of American Medical Association, British Medical Journal, and Annals of Internal Medicine) in order to assess reporting, to describe current practice and to estimate misuse of variable selection methods. Because of their bearing on variable selection [15], other factors were also examined, *i.e.*, recoding and interaction, in the primary and sensitivity analyses.

Methods

1. Inclusion criteria

Screening was performed by the first author (T.P.-L.) using the online tables of contents of the big-five medical journals for original articles published between January 2017 and December 2019 and reporting observational studies with multivariate statistical models. Articles were screened on the basis of title, abstract and full text assessment.

Only observational health studies on the human subject were considered; *i.e.,* studies including human individuals (patients, healthy volunteers, or healthcare practitioners) from whom health variables were measured without any forced intervention. Design included cross-sectional, case-control, prospective and retrospective cohort studies, and some quasi-experimental studies where adjustments are needed to correct bias, with or without a control group. Only studies reporting estimates from at least one model addressing the main study objective and requiring the selection of a set of covariates were included. Machine learning models were also considered except in cases where the variables could not be selected by a human (e.g. pixel array). Economic, genetic, descriptive epidemiological studies, meta-analyses (or pooled cohorts) and systematic reviews were excluded. Finally, only research articles having a *predictive* or *explicative* main research objective (as defined below) were included.

2. Data extraction

Types of research question

[2] The *type of research question*, in the abstract, at the end of the introduction section, in the methods section and the discussion, was categorized as either *predictive* or *explicative; explicative* questions were further divided into *causal* and *risk factor/association*. A research

question was considered as *predictive* if authors specified that they aimed to build a "predictive model" or a "prognostic model". A research question was considered as *causal explicative* in the following situations: (i) the exposure is a treatment, environment (e.g. pollution) or health behavior (e.g. tobacco consumption) that may be controlled, (ii) the authors used keywords related to causality such as 'mediation', 'causal path', 'causal association' (we did not consider reverse causality bias as it can be present in a predictive setting, or confounder as it can be used in risk factor/association studies), (iii) the authors considered that the exposure has an 'impact' or may 'affect' the outcome, (iv) use of propensity score, (v) the authors discussed or concluded that modifying the exposure might modify the outcome or suggested a policy controlling theexposure. For non modifiable variables such as race, distinction between causal or risk factor/association was made on a case-by-case basis. A research question was considered as *risk factor/association explicative* in the following situations: (i) authors used "risk factor" without further specification, (ii) none of the previous definitions (predictive or *causal explicative*) enabled to categorize the research question. No articles were simultaneously assigned to the two categories.

Variable selection methods

The variable selection method was searched in the multivariate model of each article. If there were several analyses, only the primary analysis corresponding to the primary aim was retained at first. The variable selection method was categorized according to three exclusive groups: -"Knowledge-based" including "knowledge-based without citation": the article provides bibliographic references supporting the process of selecting covariates (at least for one covariate) and " knowledge-based with citation": the article uses terms suggesting that the variables are selected based on knowledge or hypothesis prior to the analysis as suggested by "**known** to be confounders/potential confounder", "previously found to be associated" and '*a priori*', or reports that the adjustment variables have been chosen with an explicit thought of causal pathways, as well as an analysis labelled as a "mediation analysis".

- “data-driven method”: the article specifies that at least one data-driven method was used for automatically selecting covariates;

- “*unclear*”: the article is unclear, where reporting was insufficient to allow categorization into data-driven or knowledge-based methods.

In case an article used a combination of knowledge-based and data-driven methods, it was classified as using data-driven methods. The choice between alternative coding (continuous age versus age groups) and the selection of interaction terms were considered as part of the variable selection process. Hence, if authors statistically tested which recoding or interaction term induced the best model fit in order to include or not an interaction term or modify the coding of variables in the final model, the article was considered as using a data-driven method. The only exception being tests of interaction between covariates and time, usually tested in Cox regression models since they primarily serve a purpose of validating model assumptions rather than including additional terms in the model.

Studies categorized as having used “data-driven” methods, were further categorized according to 11 non-exclusive non-limitative method types defined by a prior literature search [6,7,13]: Backward elimination, Forward selection, Stepwise selection, Univariate selection, LASSO, Elasticnet, Change-in-estimate criterion, Purposeful method [16], High dimensional propensity score [17] considered as a “*data-driven method*” because the list of covariates included is data-driven, inclusion of an interaction term in the final model depending on the result of interaction test, variable coding depending on a test of linear fit.

Then, we searched variable selection in the sensitivity analyses section concerning the primary aim because information about variable selection may be managed in a sensitivity analysis. Sensitivity analyses were recorded in three non-exclusive categories depending on their objective and category “none” as follows: (i) alternative variable adjustment; (ii) alternative recoding; (iii) other sensitivity analyses. In case of data-driven alternative variable

adjustment or recoding, the name of the method was recorded. Sensitivity analyses were considered only if authors employed the term ‘sensitivity analyses’.

Excerpts

Selected excerpts were collected by the first author (T.P.-L.) when they were found to be particularly illustrative of the reporting of a method.

Source of information

Information on variable selection methods and excerpts was only collected from the methods section of each article but information on sensitivity analyses was searched in the entire article. Appendices were not considered. All references were reviewed blindly by two authors of this work: T.P.-L. and R.J. In case of disagreement, the article was read by a third author, A.L, and decisions were based on consensus or majority vote.

1. Statistical analysis

As data-driven methods could bias inference in case of causal questions they were considered as a misuse. The proportions of the variable selection methods are presented separately for the primary and sensitivity analyses. To estimate the misuse of data-driven methods, the distribution of the type of research question is described in this subgroup, including articles where data-driven methods were used in primary analyses or sensitivity analyses.

We conducted two post-hoc sensitivity analyses to explore the results about unclear reporting. First the definition of “knowledge based” was extended considering the articles that present lists of covariates right after sentences such as ‘these are (potential) confounders/mediators’. Second, 25 articles were randomly selected with variable selection methods that were still rated as “unclear” (from the method section alone, see above) after the extension of the “knowledge-based” definition. Additional information on variable selection methods (and its location) was searched in all 25 articles, appendix included. Confidence intervals at 95% were

computed with Clopper-Pearson method. Data-Management and statistical analyses were performed with R statistical software (version 3.5, The R Foundation for Statistical Computing, Vienna, Austria).

Results

2. Screening

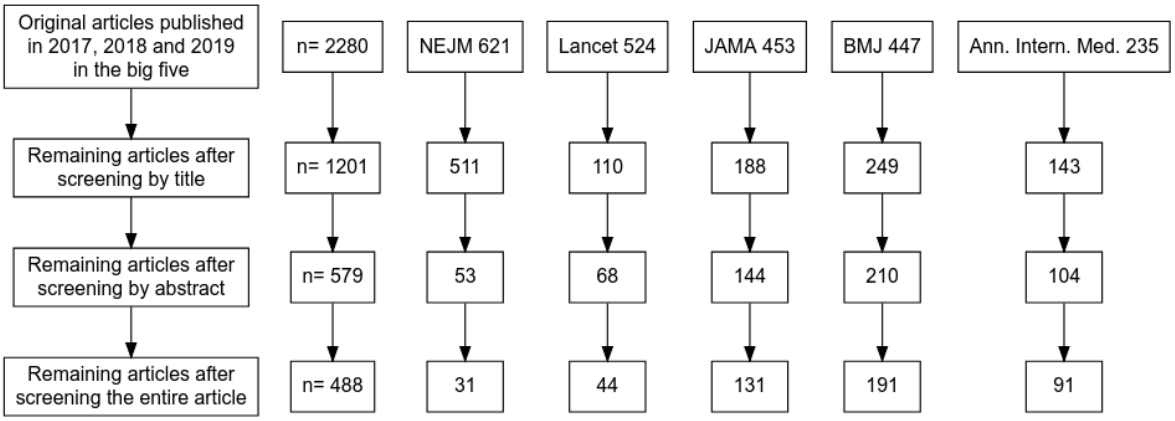


Figure 83: Flowchart review

Overall, 488 articles fulfilled inclusion criteria and were selected. Articles from the BMJ represented 39% of all included articles whereas the Lancet and the NEJM respectively

represented 9% and 6.4% (Figure 1 and supplementary figure S1 for details about excluded articles).

Practice and reporting:

The variable selection method was unclear in 234 (48%) of the 488 articles. Data-driven methods were used in 78 (16%) articles in their primary analyses and in 83 (17%) if data-driven methods were searched in both the primary and sensitivity analyses rather than in the primary analysis alone. Among the 83 articles which used data-driven methods the majority (43 articles) used test-based methods (univariate selection, backward, stepwise, forward, purposeful, double univariate analysis and unclear data-driven methods). Data-driven methods were not systematically explained 3 (3.6%). Both change-in-estimate and univariate selection methods used in 29 articles were not named properly but were explicit enough. Alternative variable adjustment and alternative recoding performed in sensitivity analyses concerned respectively 101 (20.6%) and 30 (6.2%) of the articles (Table 1).

Table 1: distribution of variable selection methods in the main analysis, then in the sensitivity analyses of 488 articles

	Total n(%)
Variable selection method in main analysis	
Unclear	234 (48%)
Knowledge-based	176 (36%)
Knowledge-based with citation	82 (16.8%)
Knowledge-based without citation	94 (19.2%)
Directed acyclic graph	9 (1.8%)
Data-driven methods (detailed category non-exclusive)	78 (16%)
Univariate selection	18 (3.7%)
Interaction test	15 (3.1%)
Linearity test	13 (2.7%)
Change-in-estimate	8 (1.6%)
Stepwise	7 (1.4%)
Backward	6 (1.2%)
High dimensional propensity score	4 (0.8%)
Elastic Net	2 (0.4%)
Forward	1 (0.2%)
LASSO	1 (0.2%)
Purposeful	1 (0.2%)
Principal Component Analysis	1 (0.2%)
Deletion, substitution, and addition [18]	1 (0.2%)
Regression tree [19]	1 (0.2%)
Net reclassification index [20]	1 (0.2%)
Kernel regularized least squares [21]	1 (0.2%)
Colinearity test	1 (0.2%)
Double univariate analysis*	1 (0.2%)
Unclear data-driven method	3 (0.6%)
Variable selection method in Sensitivity analyses (detailed category non-exclusive)	
Alternative variable adjustment	101(20.7%)
Data-driven methods	9(1.8%)
Univariate selection	4(0.9%)
Forward	1(0.2%)
High dimensional propensity score	1(0.2%)
Interaction test	2(0.4%)
Purposeful	1(0.2%)
Alternative recoding	30(6.1%)
Others	245(50.2%)
None	120(24.6%)

* select all variables that significantly correlate with both the exposure and the outcome

Reporting post-hoc sensitivity analysis

After applying a slightly different definition of knowledge-based method, the variable selection method of 176 (36.1%) articles remained unclear.

The variable selection method remained unclear in 15 (60% CI: 39%; 79%) articles out of 25, nine were knowledge-based with citation, information found in introduction (n=1), discussion(n=3), appendix (n=3) or abstract (n=2) and one was univariate selection post matching (information in discussion)

Excerpts

A total of ten excerpts were selected. One explicitly did not use data-driven methods: "...rather than deferring to statistical criteria.". Three mentioned prior knowledge: " ...existing literature…", "...a priori assumptions...", "... reviewing the literature and consulting clinical experts.". Four defined relation between variables: "... identified potential confounders…", " ...potential mediators.", "... roles as either confounders or mediators", "... risk factors may be in the causal pathway.". Data-driven methods are presented for recoding and interaction. (Table 2).

Table 2: selected excerpts from articles

	Excerpts
Variable selection	As recommended, we identified potential confounders based on existing literature, rather than deferring to statistical criteria. [22] †
Variable selection	Model 2 was the primary model because model 3 risk factors may be in the causal pathway. [23] †
Variable selection	...adjusting for such variables is known to result in reduced precision and potential amplification of bias. [24] †
Variable selection	We included covariates on the basis of a priori assumptions about their roles as either confounders or mediators. [25] †
Sensitivity analysis	In a sensitivity analysis we additionally adjusted for the following potential mediators. [26] †
Interaction	The regression model was supplemented by adding interactions of covariates one at a time and selecting the model with superior balance. [27] ‡
Interaction	Where there was statistical significance, we included the interaction term in the final model and expressed the results using the interaction. [28]‡
Recoding	We defined [X1] categories after reviewing the literature and consulting clinical experts. [29] †
Recoding	...using cubic spline models to account for possible non-linear relations with the outcome. [30]
Recoding	...Akaike information criterion had been considered as the most reliable, flexible criterion for fitting penalised splines in Cox Models. [31]‡

† categorized in knowledge-based, ‡ categorized in data-driven

Misuse of data-driven methods:

Misuse of data-driven methods as defined in the method section was found in 51 (10.5%) of the 488 articles. Methods such as machine learning, shrinkage methods or high dimensional propensity score were used to compute propensity score in 8 cases. In 5/8 cases, univariate selection was used after propensity score matching for adjusting on unbalanced covariates (standardized difference > X). Principal component analysis was used because of genetic variables comprised in the list of covariates. In 3 cases, data-driven methods were only used in sensitivity analyses. Interaction or linear testing was the only data-driven method used in respectively 7 cases and 8 other cases. Linear test concerning exposure in 5/9 cases and huge sample dataset (>100 000) in 4/9 cases. (Table 3).

Table 3: distribution of the type of research questions and distribution of variable selection methods for articles with data-driven methods in a primary or sensitivity analysis

	n (%)	
Articles with data-driven methods	83 (100%)	
Type of research question		
Predictive	14 (16.9%)	
Explicative	69 (83.1%)	
Risk/Association	18 (21.7%)	
Causal	51 (61.4%)	
Variable selection methods in the "causal" subgroup	**Primary analysis n**	**Sensitivity analysis n**
Total	48	7
Univariate selection	5	3
Interaction test	9	1
Linearity test	9	0
Change-in-estimate	8	0
Stepwise	3	0
Backward	3	0
High dimensional propensity score	4	1
Forward	0	1
LASSO	1	0
Purposeful	1	1
Principal Component Analysis	1	0
Deletion, substitution, and addition [18]	1	0
Regression tree [19]	1	0
Double univariate analysis*	1	0
Unclear data-driven method	3	0

* select all variables that significantly correlate with both the exposure and the outcome

Discussion

The percentage of unclear variable selection methods was very high, accounting for nearly 50% of all articles. However, our post-hoc analysis lowered the percentage to 36% and revealed that among articles graded as unclear 40% [21%;61%] of variable selection methods were not in the method section.The data-driven methods used are heterogeneous (test based, shrinkage, machine learning) and are not widely used (17%) in observational studies of the big five medical journals. Furthermore, more recent variable selection methods for causal inference have been published such as augmented backward elimination, or group lasso and doubly robust estimation of causal effect [32], or outcome adaptive lasso [33] but none were found in our review. The time between the publication of the method and the creation of a corresponding package, and the statisticians' habits may delay their use.

Misuse of data-driven methods was low (10.5%), corresponding to use in *causal explicative* studies, and was rare if only primary analyses were considered and interaction, linear tests were not accounted (6.8%).

Comparison with the literature

Previous reviews reported 35% of unclear variable selection methods [12,13]. There are some explanations to this difference with our review (48%). First, some associations may be considered so well known, e.g., association between smoking and cardiovascular outcomes, that authors may think they do not need to provide any explanation or citation. Second, associations could be explained and cited in other sections than the Methods section (40% in our post-hoc analysis). Third, implicit justification of selection and poor reporting may be a consequence of lack of space in the methods section but authors could add supplemental data. Fourth, the level of reporting required for an article to be categorized as "knowledge based" may impact the percentage of unclear variable selection methods (36% with a slightly different definition) but we cannot compare with previous reviews that did not detailed their “knowledge based” definition.

In previous reviews, the use of data-driven methods was more frequent: 84% for explicative and predictive studies combined in a review in two Chinese epidemiologic journals published between 2004 and 2008 [34], 35% in four major epidemiologic journals published in 2008 [12], and finally 23% in the same epidemiologic journals (explicative studies) published in 2015 [13]. For the last two reviews, stepwise selection was respectively used in 20% and 5% of the articles whereas the “change in estimate” method was used in 15% and 12% of the articles. These differences may be due to journal requirements or reporting, and publication year. However, we had a very broad definition of data-driven methods that increased their frequency compared to previous articles. Indeed, in our article interactions and recoding (even in sensitivity analyses), both data-driven were considered. Our definition of

explicative studies was dichotomized in ‘risk factor/association’ and ‘causal’ that enabled to estimate misuse of data-driven methods contrary to these previous studies.

Strengths and limitations

Only five journals were searched but these “big five” are the most widely read medical journals worldwide. The New England Journal of Medicine and the Lancet contributed few articles to this study, because most of their articles report randomised or non-randomised interventional studies.

The type of research question that we defined as *explicative causal* may be controversial because there are several definitions of causality. Moreover, we based our extraction of the *type of research question* on authors’ reporting, and our interpretation.

Variable selection methods were only screened in the methods section as reporting methods outside of the methods section is considered as poor reporting but a sensitivity analysis was carried out to measure the degree of missing information in the entire article, appendix included.

We performed an up to date review with double data extraction, making interpretation of the current practice reliable, but providing no insights on temporal trends since articles from only three consecutive years were assessed. Many authors listed covariates, with citations, or defining as confounder or potential confounder to justify their choices, but did not explicitly specify that all choices were made *a priori*, or without the use of data-driven methods. Hence, some models that we considered as being built by a knowledge-based method could have been partly built with data-driven methods, leading to an underestimation of data-driven methods in our study.

Conclusion

Stepwise and data-driven variable selection methods do not seem to be widely used in the “big five” medical journals. Unfortunately, the variable selection method is not clearly reported in many articles, and the actual proportion of data-driven variable selection may be

higher. Poor specification of the variable selection and coding scheme as well as the absence of published protocol leaves room for p-Hacking. Authors should clearly indicate that they did not use or rely on any of the data-driven methods such as those presented in the excerpts. Therefore, we recommend that the STROBE recommendation to authors to "Describe **all** statistical methods […]" be taken to the letter.

Conflict of interest statement

Declarations of interest: none

Acknowledgments

The authors are grateful to Nikki Sabourin-Gibbs, Rouen University Hospital, for her help in editing the manuscript and Emeline Lejeune for helping to extract the list of articles from the online tables of contents.

References

1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiol Camb Mass. 1999 Jan;10(1):37–48.

2. Hamaker HC. On multiple regression analysis. Stat Neerlandica. 1962 Mar;16(1):31–56.

3. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol. 1996 Jan;58(1):267–88.

4. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005 Apr;67(2):301–20.

5. Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented Backward Elimination: A Pragmatic and Purposeful Way to Develop Statistical Models. Olivier J, editor. PLoS ONE. 2014 Nov 21;9(11):e113677.

6. Desboulets L. A Review on Variable Selection in Regression Analysis. Econometrics. 2018 Nov 23;6(4):45.

7. Witte J, Didelez V. Covariate selection strategies for causal inference: Classification and comparison. Biom J Biom Z. 2019 Sep;61(5):1270–89.

8. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J Biom Z. 2018 May;60(3):431–49.

9. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Second edition. Cham Heidelberg New York: Springer; 2015. 582 p. (Springer series in statistics).

10. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med. 2007 Oct 16;147(8):W163-194.

11. Sharp MK, Bertizzolo L, Rius R, Wager E, Gómez G, Hren D. Using the STROBE statement: survey findings emphasized the role of journals in enforcing reporting guidelines. J Clin Epidemiol. 2019 Dec;116:26–35.

12. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733–6.

13. Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. Eur J Epidemiol. 2019 Aug;34(8):725–30.

14. 2016 Journal Impact Factor, Journal Citation Reports (Clarivate Analytics, 2020)

15. for TG2 of the STRATOS initiative, Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagn Progn Res. 2020 Dec;4(1):3, s41512-020-00074–3.

16. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med. 2008 Dec 16;3:17.

17. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiol Camb Mass. 2009 Jul;20(4):512–22.

18. Sinisi SE, van der Laan MJ. Deletion/substitution/addition algorithm in learning with applications in genomics. Stat Appl Genet Mol Biol. 2004;3:Article18.

19. Loh W. Classification and regression trees. WIREs Data Min Knowl Discov. 2011 Jan;1(1):14–23.

20. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008 Jan 30;27(2):157–72; discussion 207-212.

21. Hainmueller J, Hazlett C. Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. Polit Anal. 2014;22(2):143–68.

22. Fiolet T, Srour B, Sellem L, Kesse-Guyot E, Allès B, Méjean C, et al. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. BMJ. 2018 Feb 14;k322.

23. Zhong VW, Van Horn L, Cornelis MC, Wilkins JT, Ning H, Carnethon MR, et al. Associations of Dietary Cholesterol or Egg Consumption With Incident Cardiovascular Disease and Mortality. JAMA. 2019 Mar 19;321(11):1081.

24. Desai RJ, Bateman BT, Huybrechts KF, Patorno E, Hernandez-Diaz S, Park Y, et al. Risk of serious infections associated with use of immunosuppressive agents in pregnant women with autoimmune inflammatory conditions: cohort study. BMJ. 2017 Mar 6;j895.

25. Timpka S, Stuart JJ, Tanz LJ, Rimm EB, Franks PW, Rich-Edwards JW. Lifestyle in progression from hypertensive disorders of pregnancy to chronic hypertension in Nurses' Health Study II: observational cohort study. BMJ. 2017 Jul 12;j3024.

26. Nelson SM, Haig C, McConnachie A, Sattar N, Ring SM, Smith GD, et al. Maternal thyroid function and child educational attainment: prospective cohort study. BMJ. 2018 Feb 20;k452.

27. Helenius K, Longford N, Lehtonen L, Modi N, Gale C. Association of early postnatal transfer and birth outside a tertiary hospital with mortality and severe brain injury in extremely preterm infants: observational cohort study with propensity score matching BMJ 2019; 367 :l5678

28. Wallis CJD, Juvet T, Lee Y, et al. Association Between Use of Antithrombotic Medication and Hematuria-Related Complications. JAMA. 2017;318(13):1260–1271.

29. Thayakaran R, Adderley NJ, Sainsbury C, Torlinska B, Boelaert K, Šumilo D, et al. Thyroid replacement therapy, thyroid stimulating hormone concentrations, and long term health outcomes in patients with hypothyroidism: longitudinal study. BMJ. 2019 Sep 3;l4892.

30. Abrahami D, Douros A, Yin H, Yu OHY, Renoux C, Bitton A, et al. Dipeptidyl peptidase-4 inhibitors and incidence of inflammatory bowel disease among patients with type 2 diabetes: population based cohort study. BMJ. 2018 Mar 21;k872.

31. Lv Y-B, Gao X, Yin Z-X, Chen H-S, Luo J-S, Brasher MS, et al. Revisiting the association of blood pressure with mortality in oldest old people in China: community based, longitudinal prospective study. BMJ. 2018 Jun 5;k2158.

32. Koch B, Vock DM, Wolfson J. Covariate selection with group lasso and doubly robust estimation of causal effects: GLiDeR. Biometrics. 2018 Mar;74(1):8–17.

33. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. Biometrics. 2017 Dec;73(4):1111–22.

34. Liao H, Lynn HS. A survey of variable selection methods in two Chinese epidemiology journals. BMC Med Res Methodol. 2010 Dec;10(1):87.

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], Lina F. SOUALMIA[c,d]
[a] *CHU Rouen, Department of Biomedical Informatics, Rouen, France*
[b] *CHU Rouen, Department of Biostatistics, Rouen, France*
[c] *Normandie Univ, UNIROUEN, LITIS EA 4108, Rouen, France*
[d] *LIMICS U1142, Sorbonne Université, Paris, France*

Introduction

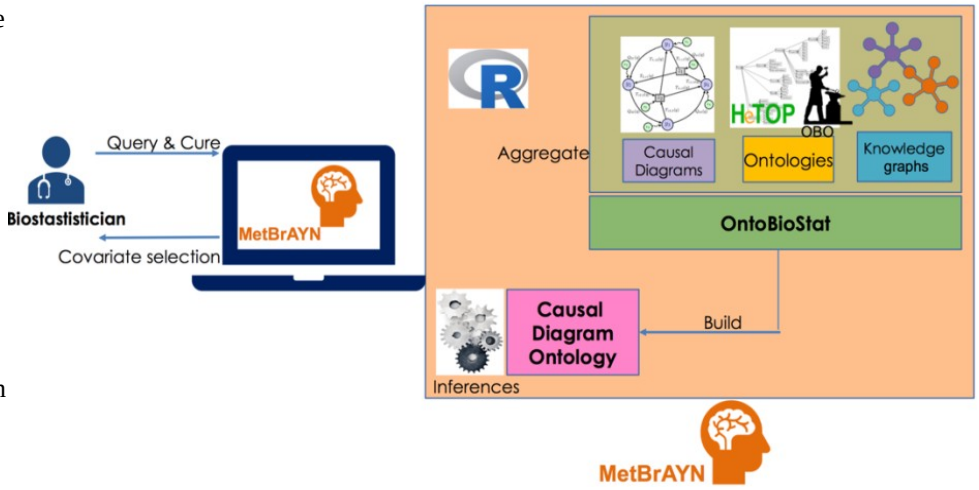
Causal diagrams (CD) [1], descendants of the directed acyclic graph (DAG), are used for covariates selection in causal inference. They have several limits as they are qualitative representations with no consensus concerning interaction, metadata inclusion or timeline information. An ontological solution called OntoBioStat (OBS) coupled to a graph database could be an answer to these limits. OBS will help biostatisticians to build causal diagrams ontology (CDO) and to obtain various sets of sufficient covariates in order to answer the research question ask.

Method

The knowledge in the ontology will be gathered from reporting guidelines of observational studies, articles or books about biostatistician consultation, DAG and CD features, and mapped with others ontologies that share similar concepts.

OntoBioStat will be built using Protégé in OWL.

In addition, a formal knowledge graph representation with a graph database will allow to enhance CD and DAG features with for example relations between relation that ontologies cannot handle. 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 [2], MetBrAYN will recommend covariates for model building, depending on the research question. The rules will be inspired by already existing methods such as back-door, front-door or disjunctive criterion. Here is presented an hypothetical use case.



Expected Results

Considering the following research question: *Do yellowed fingers cause lung cancer?*

First, the biostatistician gather and cure knowledge about the use case under the form of a graph database. The knowledge regroups literature, dataframe and study information (e.g. smoking tobacco causes lung cancer and causes yellowed fingers; smoking status is not measured in our study; Lag period of lung cancer is about one year and the exposure is yellowed fingers). Then, *MetBrAYN* relying on CDO and inference rules will provide: indispensable covariates (e.g. smoking status), optional covariates, prohibited covariates (e.g. tar deposits in lung), recoding (e.g. yellowed fingers in 5 shades), interaction term (teeth and fingers yellowed), sensitivity analysis (e.g. slight definition of yellowed fingers), with explanations (e.g. smoking is an indispensable covariates because it is a confounder), and a comment on extrapolation (e.g. extrapolation is limited because of the unmeasured confounder 'smoking status').

Conclusion

In this PhD project, the system MetBrAYN, that includes OBS and CDO, will avoid modeling mistakes and would be used to train beginners in biostatistics.

References

[1] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan;10(1):37–48.

[2] Lamy J-B, Soualmia LF. Formalization of the semantics of iconic languages: An ontology-based method and four semantic-powered applications. Knowledge-Based

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

OntoBioStat: Supporting Causal Diagram Design and Analysis

Thibaut PRESSAT LAFFOUILHÈRE[a,b,c,2], Julien GROSJEAN[a,d], Jacques BÉNICHOU[b,e], Stefan J. DARMONI[a,d], 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 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. Material & 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",

2Corresponding Author, 37 Boulevard Gambetta, 76000 Rouen, France , E-mail: t.pressat@chu-rouen.fr

"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

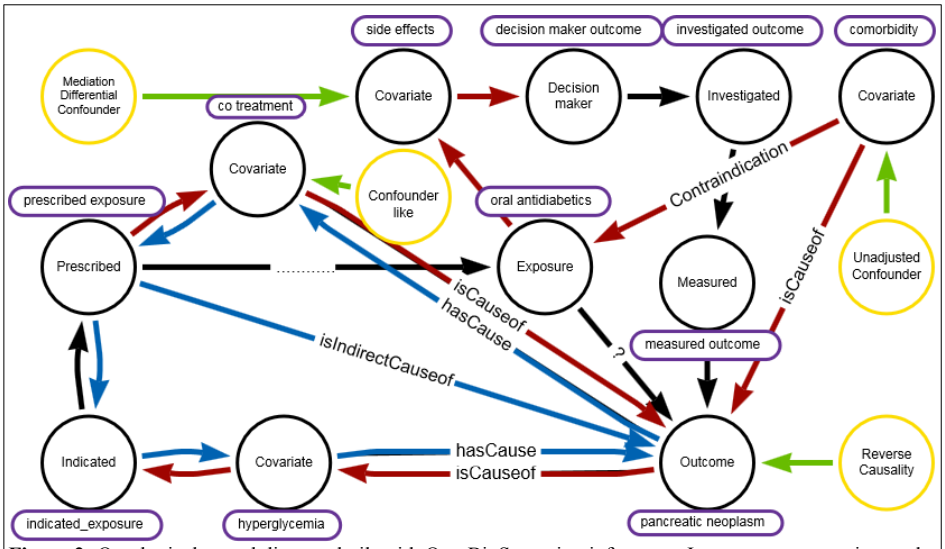


Figure 2: Ontological causal diagram built with OntoBioStat using inferences. Instances name are in purple. Object properties stated are in red, inferred in blue, being a part of automatic construction black. Inferred classes are in yellow.

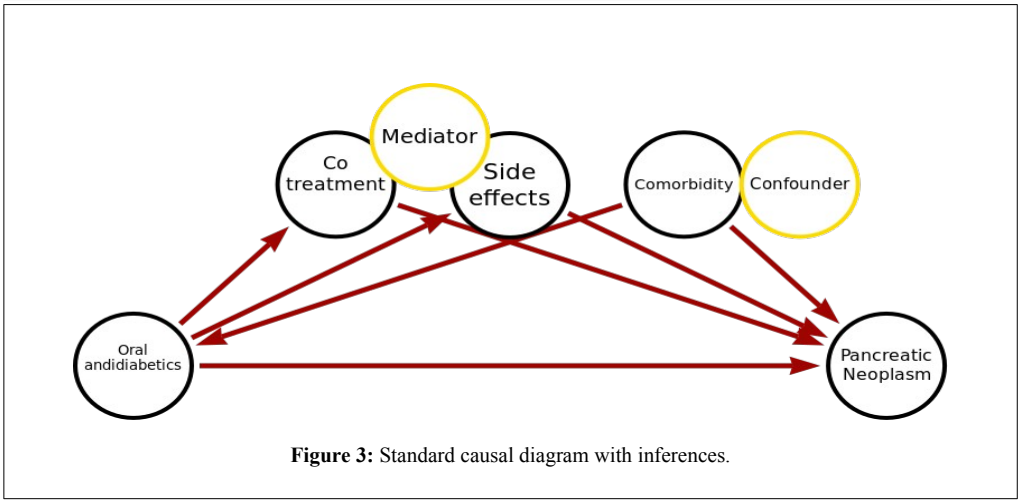


Figure 3: Standard causal diagram with inferences.

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.

4. Discussion and Conclusion

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.

Ontological Representation of Causal Relations for a Deep Understanding of Associations between Variables in Epidemiology

Thibaut Pressat Laffouilhère[1,2,3,*], Julien Grosjean[1,4], Jean Pinson[5], Stéfan J. Darmoni[1,4], Emilie Leveque[2], Emilie Lanoy[6], Jacques Bénichou[2,7], Lina F. Soualmia[3,4]

[1] CHU Rouen, Department of Biomedical Informatics, Rouen University Hospital, France
[2] CHU Rouen, Department of Biostatistics, Rouen University Hospital, France
[3]Normandie Univ, UNIROUEN, LITIS-TIBS EA 4108, Rouen, France
[4] INSERM U1142 LIMICS, Sorbonne Université, Paris, France
[5]CHU Rouen, Department of Surgery, Rouen, France
[6]Biostatistics and Epidemiology Unit, Gustave-Roussy, Villejuif, France
[7]INSERM U1018, CESP, Université Paris-Saclay, Paris, France
* Corresponding author
{t.pressat,julien.grosjean ,jean.pinson, stefan.darmoni,emilie.leveque, jacques.benichou, lina.soualmia}@chu-rouen.fr, emilie.lanoy@wanadoo.fr

Abstract. Understanding statistical results is crucial in order to spread right conclusions. In observational studies, statistical results are often reported as associations without going further. However, each association comes from causal relations. Causal diagrams are visual representations enabling to understand causal mechanisms behind the association found. In the era of big data and growing number of variables, visual approaches become inefficient. Ontological representation of causality and reasoning could help to explain statistical results. OntoBioStat is a domain ontology related to covariate selection and bias for biostatistician users. It was designed using expert corpus from comprehensive literature review, and validated by three biostatisticians accustomed to causal diagrams. In this paper, we focused on the presentation of an OntoBioStat's feature able to infer explanations about statistical associations. The ontologization of the feature of interest resulted in 14 object properties, three classes and five Semantic Web Rule Language rules. Each rule allows to infer a different object-property that explains statistical association between two variables. Rules are based on *isCauseof* statements between different individuals. OntoBioStat feature performances were illustrated through a real-life retrospective observational study. From 28 instances and 48 object properties stated, a set of 1,939 object properties were inferred. OntoBioStat explained 65% of the 48 statistical associations found. In conclusion, OntoBioStat could help to explain a part of the significant statistical associations between two variables but cannot yet predict significant ones.

Keywords: Ontology, Epidemiology, Causality

Introduction

The representation of causal relations between variables in epidemiology is crucial for estimating a true causal effect and also for explaining spurious associations between two variables. Several knowledge representations could be used such as the Sufficient Component Cause (SCC) model [1], Directed Acyclic Graphs (DAGs) [2], or Causal Diagrams (CDs) [3]. These latter enable researchers to represent variables as nodes, and causal relations as directed ($\rightarrow$), bidirected ($\leftrightarrow$) or non-directed (–) edges. A unidirectional edge from a node X to a node Y means that X is one of the causes of Y. Two variables share an ancestor if they have a common cause (bidirectional edge). Two variables share a descendant if they have a common constant (not variable) effect (non-directional edge). CDs have been widely used to represent collider-bias or overadjustment bias, both resulting in spurious associations [4,5]. Interpreting statistical association in observational studies is an open discussion in the epidemiologists' community. Researchers might be discouraged to explain all the associations (spurious or not) found during the research process, because of the amount of growing variables and

statistical tests. Moreover, young ones could misinterpret association in bivariate or multivariate analysis. CD or DAG that are actual standard for covariate selection have poor expressivity and hence limited knowledge representation.

In order to help researchers in (re)producing, understanding their statistical results and sharing hypothesis based on knowledge representation in biomedical research, ontologies could be used. Several ontologies in the domain of biology and statistics have been created such as Ontology of Biological and Clinical Statistics (OBCS) [6] mapped with Statistical Ontology (STATO) (http://stato-ontology.org) and Ontology of Biomedical Investigations (OBI) [7] or the Ontology of Clinical Research [8]. However, they do not cover the understanding of the association found between two variables. In this context, an ontology dedicated to the covariate selection and causal representations may be more useful. Statistical Learning Ontology can answer to “What are the variables correlated with ...?” but the answer is based on statements only queried with SPARQL, and not by inferences [9]. Ontologies representing causal relation are numerous [10,11,12] but their design was not driven by the competency question cited above and their formalism not adapted to the sufficient, counterfactual and necessary cause representation.

We designed the OntoBioStat [13] ontology in order to assist biostatiscians users and researchers in many tasks such as causal diagram design, covariate selection, and for providing explanations for the statistical association between two variables. In this paper, we focus on the latter feature corresponding to the seventh and last competency question. Hence, only a subset of the ontology will be presented. The paper is organized as follows: first the structure of OntoBioStat is introduced, then the feature of interest, and finally a use case is provided to illustrate our work. The obtained results are discussed, and we give an outlook to further work.

Materials and Methods

The OntoBioStat Ontology and Feature Focus

OntoBioStat is a domain ontology related to covariate selection and bias for causal inference in observational studies intended for biostatistician users. The last version is publicly available at https://bioportal.bioontology.org/ontologies/OBS. OntoBioStat is expressed using the Ontology Web Language (OWL) standard [14] and was designed using Protégé 5.5 [15]. OntoBioStat is composed by: (i) 53 classes, such as ‘Variable’, ‘Covariate’ or ‘Path_Modifier’; (ii) 33 object-properties, such as *isCauseof* or *Share_ancestor*; (iii) 11 data-properties; (iv) nine equivalent class axioms; (v) 28 instances; and (vi) 29 rules in Semantic Web Rule Language (SWRL). The reasoning process is supported by Pellet inference tool [16]. OntoBioStat provides a broad framework for knowledge representation needed in case of covariate selection for true causal estimation between two variables. Variables representation integrates ‘Meta_Variable’ subclasses such as ‘Location’ and ‘Period’. Three types of causal relation are represented: (i) counterfactual probabilistic cause represented with object properties presented in this article (e.g. *isCauseof*); (ii) necessary causes with upper classes; and (iii) sufficient cause with object properties not described here (*Contraindication* and *Absolute_indication*). Object properties related to interaction (*isModifiedby*), missing data cause (*isCauseofNA*), and the explicit statement of no causal relation (*NotCauseof*) were defined but not developed here either.

The OntoBioStat feature was built using a corpus from comprehensive literature review. Articles about covariate selection or bias for statistical modelling using DAGs (or CDs) were included. The following query was processed in PubMed: [Directed Acyclic Graph OR Causal Diagram] with English language filter. The extracted terms related to this feature are

definitions, representations, and synonyms of relations used in DAGs and CDs. The extracted corpus contains seven terms describing four relations between two variables: (i) direct effect/child/descendant is represented with a unidirectional/single headed arrow/arc/edge; (ii) common/share cause/ancestor/parent is represented with bidirectional/two-headed arrow; (iii) share descendant is represented with undirected/non-directional path (one or more succession of edges); and (iv) indirect effect (effect mediated by another variable) [3,4,17-21]. This corpus was validated manually by three biostatisticians (TPL, ELa, ELe) accustomed to the use of DAGs and observational studies. The feature's ontologization was driven by (i) corpus, (ii) tutorial for building DAG and CD published in peer review articles or books [22-24], (iii) theoretical cases from CD representation such as bidirected, non-directed, directed edges, and (iv) the seventh competency question: 'What type of relation exists between two variables?'. Indeed, OntoBioStat design was built in order to answer seven competency question about confounding and bias.

In this section we focus on 14 non-reflexive object-properties (Figure 1), three classes, and five SWRL rules (rule (1), rule (2), rule (3), rule (4), and rule (5)).

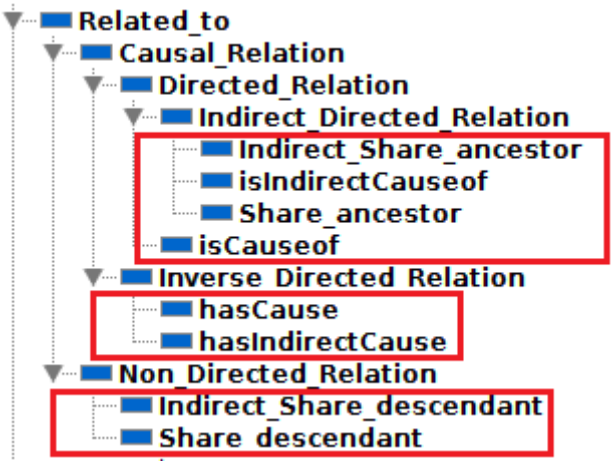


Fig. 1. OntoBioStat object properties representing variables relations.

Related_to (X, Y) is a symmetric property that means that a significant statistical association between X and Y should exist. The *Inverse_Directed_Relation* properties are inferred thanks to inverse properties. Bidirectional edges (share ancestor) and non-directional edges (share descendant) in CDs were defined as symmetric properties. *isCauseof*, *hasCause*, and *isIndirectCauseof* are asymmetric object-properties. *Causal_Relation*, *Directed_Relation*, *Indirect_Directed_Relation*, *Inverse_Directed_Relation*, *Non_Directed_Relation* were created in order to group the eight initial properties represented in Figure 1 (red framed) and in Figure 2. For a better understanding, and because the inferred properties are numerous, only few of them are displayed in the Figure 2. Indirect relations are causal relations between two variables when a third stands between them. Directed relations refer to CD representation with directed edges. Non directed relations refer to the undirected edge from CD.

Inferred_Variable (?y) ∧ Inferred_Variable(?x) ∧ *Inverse_Directed_Relation(?y, ?x)* ∧ *Inverse_Directed_Relation(?z, ?y)* → *isIndirectCauseof(?x, ?z)* (1)

Covariate(?z) ∧ Inferred_Variable(?y) ∧ *isCauseof(?z, ?x)* ∧ *isCauseof(?z, ?y)* → *Share_ancestor(?x, ?y)* (2)

Path_Modifier(?z) ∧ Inferred_Variable (?x) ∧ Inferred_Variable(?y) ∧ *isCauseof(?y, ?z)* ∧ *isCauseof(?x, ?z)* → *Share_descendant(?x, ?y)* (3)

Covariate(?y) ∧ *Share_ancestor(?x, ?y)* ∧ *Inverse_Directed_Relation(?z, ?y)* → *Indirect_Share_ancestor(?x, ?z)* (4)

Path_Modifier(?z) ∧ Inferred_Variable(?y) ∧ Inferred_Variable(?x) ∧ *Directed_Relation(?y, ?z)* ∧ *Indirect_Directed_Relation(?x, ?z)* → *Indirect_Share_descendant(?x, ?y)* ∧ *hasDescendant(?y, ?z)* ∧ *hasDescendant(?x, ?z)*

(5)

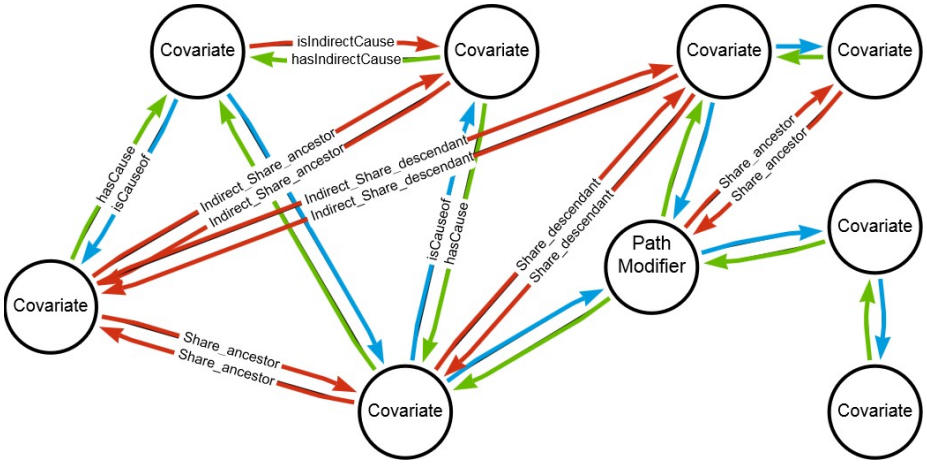


Fig. 2. Example of inferences based on *isCauseof* statements (in blue). Inverse properties (in green) and partial inferences are based on OntoBioStat SWRL rules (in red).

Grouping the eight properties in upper object properties enables to reduce the number of SWRL rules. Indeed, SWRL rules do not allow the use of 'OR'. For example, *Share_ancestor* is defined by: *isCauseof* and *isCauseof* or *isCauseof* and *isIndirectCauseof* or *isIndirectCauseof* and *isIndirectCauseof*. *Share_ancestor* is also summarized by *Inverse_Directed_Relation* and *Inverse_Directed_Relation* (rule 1). *Related_to* descendants are inferred based on *isCauseof* statements.

The class 'Covariate' *subClassof* 'Variable' has to be stated in order to infer the *Causal_Relation* descendants. Concerning the *Non_Directed_Relation* descendant, the class 'Path_Modifier' has to be stated instead of 'Covariate'. A path modifier is a variable that is constant and adjusted for or use in matching methods in order to estimate true causal effect between two variables [25]. The class 'Inferred_Variable' groups 'Covariate', 'Exposure', and 'Outcome'. These three classes are useful in order to answer different competency questions about confounding that are not developed here. 'Path_Modifier' is a "dead end". Hence, inferred edges may be pointed at it but cannot start from, it except in case of *Share_ancestor* or *hasCause* (Figure 2).

In all the rules *x*, *y*, and *z* are different individuals in order to respect the non-reflexive object properties. Indeed, given four variables *w*, *x*, *y*, *z*: *z isCauseof x* and *w*, and the couple (*x*, *w*) *isCauseof y*. Without 'differentFrom' statements *y Indirect_Share_ancestor* with *y*. In the rule (1) *x* and *y* must be 'Inferred_Variable' because transitivity is interrupt by any 'Path_Modifier'. In the rule (4) *y* must be a 'Covariate' for the same reason. In the rule (5) *Directed_Relation* (→, → →, ↔, ↔ →) and *Indirect_Directed_Relation* (→ →, ↔, ↔ →) enable to infer two types of *Indirect_Share_descendant*. Indeed, two variables indirectly share a descendant if one of them is a direct cause of descendant, or if both are indirect cause of descendant (share ancestor included) (Figure 2).

Use Case

A real retrospective observational study was used to illustrate the OntoBioStat ability to answer: “What type of relation exists between two variables?”. Using the Protégé ontology editor, based on medical knowledge about digestive surgery, the users (TPL helped by JP) entered: (i) the name of the instances that are measured and unmeasured variables (‘Covariate’ in our case), and path modifiers (Adjustment_Covariate, Matched_Covariate, Stratification_Covariate); then (ii) the causal relations between variables (*isCauseof* object property only) as if a causal diagram or a causal directed acyclic graph was drawn. Finally, the Pellet reasoner was activated. In this use case, some OntoBioStat features were not used such as those based on the classes ‘Necessary_Variable’ or ‘Meta_Variable’ which increase the number of object properties inferred.
This use case is based on a real dataset about digestive surgical procedures. A subset of 16 variables and 102 observations with no missing data were used. Observations from the subset were increased from 102 to 300 observations using sampling with replacement in order to reach sufficient statistical power and hence enable more discussion about indirect causal relation. Associations between each pair of variables were assessed using Pearson correlation and plotted. The association between two variables was expressed as *p-value* (no estimate and no confidence interval). Significant *p-values* (<0.05) were explained with OntoBioStat inferences. The Protégé interface provides ‘explain inference’ feature that displays the reasoner steps for a given inference. Object properties are not mutually exclusives. For example, an association could be explained by *Share_ancestor* and *isCauseof*. Two multivariate models were computed in order to give examples of the class ‘Path_Modifier’ impact. The models results were expressed with estimate and *p-value* before and after adjustment. Statistical analyses were performed using R software and the corrplot package.

Results

An amount of 28 different instances of ‘Covariate’ and 48 *isCauseof* were entered, and then 1,939 object properties were inferred in less than five seconds (Figure 3).
A total of 48 p-values were significant (Figure 4) and 31 were explained by the following object properties : (i) ten *isCauseof* (e.g. **corticosteroid therapy** *isCauseof* **immunosuppression**); (ii) 12 *isIndirectCauseof* (e.g. **sex** *isIndirectCauseof* **ASA score** because **sex** *isCauseof* **arthritic disease**, **arthritic disease** *isCauseof* **organe failure at the beginning** *isCauseof* **ASA score**) ; (iii) ten *Share_ancestor* (e.g. **bmi** and **diabete** *Share_ancestor* **health behavior**) ; (iv) 11 *Indirect_Share_ancestor* (e.g. **surgical complication at 90 days** (Clavien-Dindo) *Indirect_Share_ancestor* with **post operative intensive care unit** because Hinchey *isCauseof* **organe failure at the beginning** and **surgical complication at 90 days** (Clavien-Dindo) and **organe failure at the beginning** *isCauseof* **post operative intensive care unit**). Two explanations (two different object properties) were given for 12 associations.

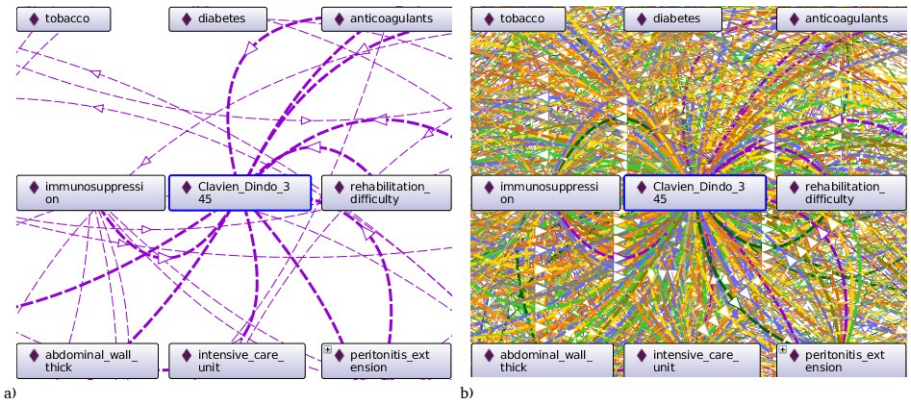


Fig. 3. a) isCauseof object properties stated, b) object properties inferred with OntoBioStat.

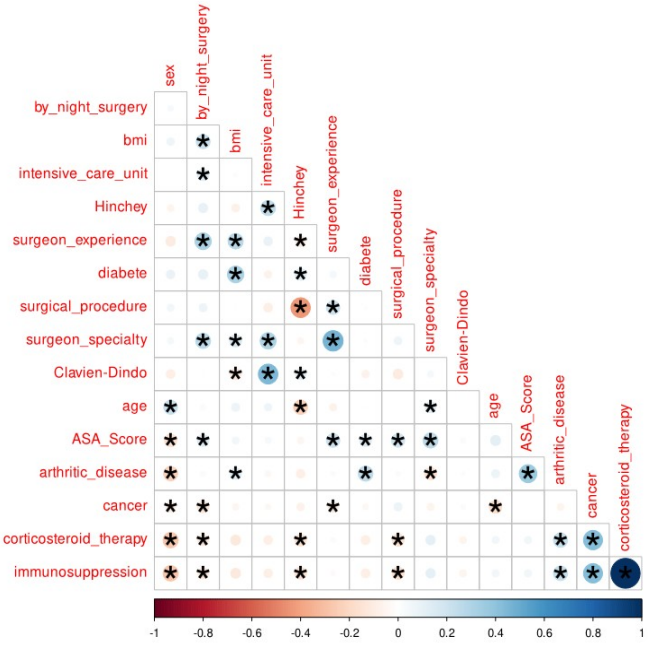


Fig. 4. Correlation matrix of 16 variables. Significant correlations ($p<0.05$) are highlighted with a star symbol.

The first multivariate statistical model is a linear regression with **ASA score** as outcome and **sex** and **arthritic disease** as explicative variables. Before adjustment **sex** and **ASA score** are associated because **sex** *isIndirectCauseof* **ASA score** (*p-value* = 0.004, estimate = -0.25). After adjustment for **arthritic disease, sex** is no longer associated with **ASA score** because the path was intercepted by the 'Path_Modifier' **arthritic disease** (*p-value* = 0.09, estimate = -0.14). The second multivariate statistical model is a logistic regression with **immunosuppression** as outcome and **surgeon specialty** and **by night surgery** as explicative variables. Before adjustment **surgeon specialty** and **immunosuppression** are not associated (*p-value* = 0.319, estimate = 0.72). After adjustment for **by night surgery, immunosuppression** is nearly associated with **surgeon specialty** because they *Share_descendant* **by night surgery** (*p-value* = 0.08, estimate = 1.36).

Discussion

In this paper, we presented OntoBioStat a new causal representation and one of its features: a novel approach for explaining statistical associations in observational studies based on SWRL rules.

Concerning causal representation, OntoBioStat succeeds in the implementation of the CDs edge representations (non-directional, bidirectional, and unidirectional). Object properties were created based on a corpus of epidemiological articles about CDs or DAGs, hence the vocabulary is understandable by CDs and DAGs users. Object properties respect mathematical logic which defines association based on causal reasoning. Indeed, two variables are associated if: (i) one of the two causes the other, (ii) they share an ancestor, and (iii) they have a common descendant which is constant in the sample studied. This causal representation is basic compared to other ontologies of causation representing *causal-like* properties such as *allow* and distinguishing 'states' from 'event' classes [10]. *isCauseof* might be seen as a top object property for object properties from other ontologies such as *directly negatively regulates activity of* from OBO Relation Ontology that gather various relations from the Open Biological and Biomedical Ontology [26]. Existing ontologies that represent causation, such as [27] and [10], inferred *explain* or *maycause* object properties using transitivity and is_a relation. However, they do not represent indirect relation such as *Share_ancestor* or sufficient cause such as *Contraindication*, and cannot infer any common descendant. The class 'Path_Modifier' integration provides a significant added value to OntoBioStat representation, and hence reasoning compared to other causal ontologies. Path Modifier class enables to deal with variables that are constant or adjusted for, in order to infer *Share_descendant* object property and 'block' some inferences.

Concerning the feature itself, previous tools developed such as R package dagitty [28] provide many features (path analysis, minimal sufficient adjustment sets) but not the explanation feature nor the complex causal representation.

Concerning statistical association explanations, unlike previous manual and visual approaches using DAG or CD representations, OntoBioStat provides automatic inferences. Indeed, visual approaches become inefficient with the growing number of variables. In our use case, OntoBioStat explained 64.5% of the associations, and 35.5% of the associations remained unexplained. These could be due to erroneous initial knowledge (wrong statements) and multiple statistical testing. Indeed, 136 statistical tests were computed hence near than five percent (n=7) of these tests discovered an association by mistake. Moreover, a *p-value* <0.05 does not necessarily mean that an association is clinically significant. Object properties inferred are numerous (n=1,939) but not always useful for the biostatistician and redundant. Indeed, *isCauseof* statements produce the five following inferences because of hierarchy and inverse relations: *hasCause*, *Inverse_Directed_Relation*, *Directed_Relation*, *Causal_Relation*, and *Related_to*. These five are not always relevant but are needed for simpler SWRL rules and deliver superficial information. In a final interface, inferences would be provided without irrelevant information and focus by default on the eight object properties framed in red in Figure 1.

In conclusion, despite this interesting 'explanation' feature, it is important to note that OntoBioStat cannot predict significant associations, neither their strength nor their sign. In order to improve OntoBioStat formalism, *isCauseof* object property will be mapped with other ontologies such as the OBO Relation Ontology. In a future use case, OntoBioStat domain will be extended from covariate selection and bias in observational study to knowledge mining using existing drug knowledge graph with thousands of relations [29].

References

1. Rothman KJ, Greenland S. Causation and Causal Inference in Epidemiology. Am J Public Health. 2005 Jul;95(S1):S144–50
2. Pearl J. Causality: models, reasoning, and inference. Cambridge, U.K. ; New York: Cambridge University Press; 2000. 384 p
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan;10(1):37–48.
4. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009 Jul;20(4):488–95.
5. Westreich D. Berkson's bias, selection bias, and missing data. Epidemiology. 2012 Jan;23(1):159–64.
6. Zheng J, Harris MR, Masci AM, Lin Y, Hero A, Smith B, et al. The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis. J Biomed Semant. 2016 Dec;7(1):53.
7. Bandrowski A, et al. The Ontology for Biomedical Investigations. PLoS ONE. 2016 Apr 29;11(4):e0154556.
8. Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of Clinical Research (OCRe): An informatics foundation for the science of clinical research. Journal of Biomedical Informatics. 2014 Dec;52:78–91.
9. Behnaz A, Bandara M, Rabhi FA, Peat M. A Statistical Learning Ontology for Managing Analytics Knowledge. In: Mehandjiev N, Saadouni B, editors. Enterprise Applications, Markets and Services in the Finance Industry p. 180–94. (Lecture Notes in Business Information Processing; vol. 345).
10. 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.
11. 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.
12. Rovetto, Robert John and Riichiro Mizoguchi. Causality and the ontology of disease. Appl. Ontology 10 (2015): 79-105.
13. Pressat Laffouilhère T, Grosjean J, Bénichou J, Darmoni SJ, Soualmia LF. Ontological Models Supporting Covariates Selection in Observational Studies. Stud Health Technol Inform. 2021 May 27;281:1095–6.
14. Bock, A. Fokoue, P. Haase, R. Hoekstra, I. Horrocks, A. Ruttenberg, U. Sattler, M. Smith, OWL 2 Web Ontology Language, W3C recommendation
15. 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.
16. Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y. Pellet: A practical OWL-DL reasoner. Web Semantics. 2007 Jun;5(2):51-53.
17. Howards PP, Schisterman EF, Poole C, Kaufman JS, Weinberg CR. 'Toward a clearer definition of confounding' revisited with directed acyclic graphs. Am J Epidemiol. 2012 Sep 15;176(6):506–11.
18. VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. Am J Epidemiol. 2007 Nov 1;166(9):1096–104.
19. VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology. 2007 Sep;18(5):561–8.
20. VanderWeele TJ. Mediation and mechanism. Eur J Epidemiol. 2009;24(5):217–24.
21. Shpitser I, VanderWeele TJ. A complete graphical criterion for the adjustment formula in mediation analysis. Int J Biostat. 2011 Mar 4;7(1):16.
22. Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol. 2021 Aug 8;S0895-4356(21)00240-7.
23. VanderWeele TJ, Staudt N. Causal diagrams for empirical legal research: a methodology for identifying causation, avoiding bias and interpreting results. Law Probab Risk. 2011;10(4):329–54.
24. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol. 2008 Oct 30;8:70.
25. Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet. 2002 Jan;359(9302):248–52.
26. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46.
27. Besnard Ph, Cordier M-O, Moinard Y. Ontology-Based Inference for Causal Explanation. In: Zhang Z, Siekmann J, editors. Knowledge Science, Engineering and Management. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 153–64. (Lecture Notes in Computer Science; vol. 4798)
28. Ankan A, Wortel IMN, Textor J. Testing Graphical Causal Models Using the R Package "dagitty". Current Protocols. 2021 Feb.
29. Lelong R, Dahamna B, Leguillon R, Grosjean J, Letord C, Darmoni SJ, Soualmia LF. Assisting Data Retrieval with a Drug Knowledge Graph. Stud Health Technol Inform. 2022 Jan 14;289:260-263

OntoBioStat: an ontology for covariate selection within the framework of causal inference in observational studies

Thibaut Pressat Laffouilhère[1,2,*], Julien Grosjean[3,4], Jean Pinson[5], Stefan J. Darmoni[3,4], Émilie Levêque[7], Émilie Lanoy[6], Jacques Benichou[7,8], Lina F. Soualmia[2,4]

[1] *Clinique Ambroise Paré, groupe ELSAN Department of medical information, 387 Rte de Saint-Simon, F-31100 Toulouse, France.*
[2] *Univ. Rouen Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen*
[3] *CHU Rouen, Department of Biomedical Informatics, Rouen University Hospital, France*
[4] LIMICS, INSERM U1142, Sorbonne Université & Université Sorbonne Paris Nord, Paris, France
[5] CHU Rouen, Department of Surgery, Rouen, France
[6] Institut Pierre Louis d'Épidémiologie et de Santé Publique (IPLESP), INSERM, Sorbonne Université, Paris, France
[7] CHU Rouen, Department of Biostatistics, Rouen University Hospital, France
[8] CESP, INSERM U1018, Université Paris-Saclay, Paris, France
**pressat@elsan.care, Clinique Ambroise Paré, groupe ELSAN Department of medical information, 387 Rte de Saint-Simon, F-31100 Toulouse, France*

Abstract
Variable selection is a key step in statistical modeling of observational epidemiologic data when the goal is to assess a causal link between an exposure and an outcome. Causal diagrams are a means to appropriately select the relevant covariates but they suffer from several limitations. For example, it is not possible to represent a necessary cause, nor the shape of the causal relation. Furthermore, causal diagrams construction lacks formalism and implicit knowledge is not represented. To cope with these limitations and to fill the gap in knowledge representation for covariate selection in causal inference models developed from observational studies, the proposed approach is to rely on the use of an ontology. This paper presents the OntoBioStat ontology and its use through a real-world observational study. The ontology design was driven by eight competency questions, each one addressing an issue of covariate selection or potential bias in the study, a literature review about directed acyclic graphs, reporting guide-lines and previous ontologies. Using OntoBioStat, we represented a real retrospective observational study that compares two surgical procedures. From these data, the automatic reasoning process based on axioms and a set of rules from the Semantic Web Rules Language allowed us to infer 240 classes and 1,465 object properties from 28 instances (i.e., variables) and 49 object properties. OntoBioStat inferences also pointed out 12 instances that should be adjusted for and three instances that could be used as proxies for an unmeasured confounder. These results show that OntoBioStat could be used as a complementary resource in decision-support systems for covariate selection enhanced with explanations.

Keywords: Covariate selection, Causality, Ontology, Ontology Web Language, Bias, Observational study, Decision support techniques.

Statement of significance:
Problem: Covariate selection using causal diagrams has important limitations for causal inference and needs improvement
What is Already Known: Causal diagrams are underused, due to their lack of formalism and poor knowledge representation. While there are alternative forms of knowledge representation for causal inference, none seem particularly suitable for covariate selection.
What This Paper Adds: The development and use of an ontology, namely OntoBioStat, provides a useful framework for causal diagram construction by making explicit use of implicit medical knowledge, overcomes the limits of the causal diagram knowledge

representation, determines potential covariates for inclusion in statistical models, and also explains why they should be included.

1. Introduction

Randomized controlled trials have been historically considered as the gold standard for addressing causality questions in clinical research. Randomization aims at balancing known and unknown patient characteristics between treatment arms, hence controlling for confounding and yielding unbiased estimation of treatment effects [1]. C*onfounders* [2] also called *lurking* variables [3] are defined as covariates (e.g., hemorrhage) that cause directly or indirectly the outcome (e.g., death), and that are associated with the exposure (e.g., hemorrhage treatment) without being caused by it (e.g., hemorrhage is not caused by treatment). Despite their strengths, randomized controlled trials cannot be conducted for all exposures, especially when exposure is not a drug treatment (e.g., upon assessing the effect of smoking or alcohol consumption on otorhinolaryngology cancer), for obvious ethical and practical reasons [4].

In observational studies (e.g., cohort or case control studies), causal effect estimation suffers from systematic multiple biases usually grouped in three categories: confounding, information bias, and selection bias [5]. These biases may result in a spurious relation between exposure and outcome or conversely may mask an existing relation [6]. Regression models (e.g., ordinary least squares linear regression) have been widely used to adjust for multiple confounders. However, while adjusting for true confounders will contribute to unbiased estimation of causal effects, adjusting for mediators or colliders will incur on the contrary extra bias and must be avoided. A mediator is an intermediate covariate between the exposure and the outcome [7] (Fig 1). A collider on a path is a common consequence of at least two variables of this path (e.g., exposure and outcome) [8]. It is thus important to be able to distinguish mediators and colliders from confounders. Causal diagrams (CDs), a type of directed acyclic graphs (DAGs), are graphical displays in which variables are represented as

nodes, and their causal relation as edges (directed or not). CDs enable to visualize the variables involved in a specific causal question and help in selecting the relevant set of covariates to be included in statistical models [9]. Several algorithms have been proposed to assist in covariate selection such as the back-door criterion [10], or the disjunctive criterion [11]. Some user friendly-tools have been developed in order to draw causal DAGs and select a minimal set of covariates, such as dagitty, a web interface and R package [12].

Despite CD usefulness for knowledge representation and the availability of accessible tools such as dagitty, CDs remain underused by the biostatistics users. For instance, a literature review highlighted that CDs had been used in only 1.8% of the observational studies published in five major medical journals in the years 2017-2019 [13]. Another recent review of articles using CDs reported that only two thirds of CD users shared their CD, and barely half of the articles reported the set of adjustment covariables obtained from CDs [14]. Biostatistics users need accessible and complete guidelines for creating CDs [15]. The existing methods or recommendations for building CDs (i) are based on general principles (e.g., the absence of edge between two variables implies that there is no causal link), (ii) only provide basic recommendations to prevent covariate selection errors such as selecting colliders [16], or (iii) describe knowledge-based building methods potentially not applicable for a particular study setting or research question [17]. Furthermore, CDs only use limited knowledge representation and do not contain all knowledge required for covariate selection. For instance, in the diagrams, an edge signifies a counterfactual probabilistic causal relation between two variables, while it might be beneficial to depict necessary, sufficient, or deterministic causal relations. Other ways of representing causality exist, such as Granger causality and Dynamic Bayesian networks [18]. However, they suffer from the same limitations as CDs in terms of knowledge representation and are not commonly used by biostatisticians in their studies.

In this context, an application ontology related to the covariate selection domain could be a promising approach to help in listing and standardizing the knowledge needed to design and build CDs. Moreover, it may promote the storage, sharing, reuse and enhancement of CDs. Indeed, ontologies [19] may be used as data models with an explicit framework for knowledge representations. Ontologies are composed of classes (concepts), object properties (roles) and data properties that may be articulated with Semantic Web Rule Language (SWRL) rules and axioms, allowing to make explicit inferences from implicit knowledge. Classes are comparable to categorize or types of nodes that allow to categorize variables of a dataset (called instances in the ontology framework). For example, the variable 'location of the study' is categorized in the class **Location**. Object properties are similar to directed edges linking two variables (instances) but they may represent something else than a causal link (e.g., A *hasLocation* B instead of A *causes* B). Dataproperties correspond to additional information about a given instance such as the proportion of missing data. Axioms allow to define equivalence and subsumption rules between two classes (nodes) or two object properties (directed edges). SWRL rules allow to make inferences that axioms cannot solve. That is why ontology belongs to symbolic artificial intelligences, natively “explainable” unlike deep learning, machine learning or statistical learning methods, which are black boxes. Ontologies provide transparent reasoning, and may be used in decision-support systems. Moreover, as graph representation, ontologies can be queried using SPARQL or other query languages for graphs.

Some ontologies related to biomedical research have been designed to document statistical procedures from data collection to conclusion [20-22]. Other relate to causality issues, such as the Radiology Gamut Ontology in radiology [23] or the Genetic Ontology Causal Activity Model (GO-CAM) in genetics [24] and some generic ontologies finely represent different causal like relation such as Ontology of Causal Relation with object properties named: *allow*, *perpetuate* and *maintain* [25] As far as we know, only two existing

ontologies are related to the variable selection process: (i) Statistical Learning Ontology (SLO) which is based on statements made by the user (i.e., no use of reasoning but, only SPARQL queries) about links types between variables (e.g., causal relation or hypothetical relation) and has been used in marketing [26]; and (ii) Dataset Characteristics and Quality Ontology (DCQO) whose aim is to support feature selection for machine learning models, hence is not dedicated to represent prior causal knowledge but dataset quality (e.g., **Correlation** or **Conciseness**) [27].

This paper presents the design and modeling choices of OntoBioStat, an application ontology which aims at representing all the knowledge that may be used in covariate selection to address a causality issue in medicine regardless of the research question. Then, a detailed use-case illustrating the use of OntoBioStat as a resource knowledge in decision-support is presented.

2. Materials and Methods

Design and Modelling

OntoBioStat was built using a modified version of the method and following steps proposed by Natalya F. Noy and Deborah L. McGuinness in Ontology Development 101: A Guide to Creating Your First Ontology [28]: (i) define domain and scope: purpose and applicability are established, (ii) reusing previous ontology, (iii) enumerate important terms: comprehensive vocabulary vital for the ontology are meticulously curated, drawn from authoritative sources, and (iv) create class, object properties and data properties: in this article, the ontology was constructed incrementally, starting with fundamental DAG components and inferences, then expanded to incorporate more sophisticated variations of directed acyclic graph (DAG) structures (e.g., interaction, signed edges) and finally original ontology-specific additions.

Domain and Scope

OntoBioStat is an application ontology for covariates selection for causal inference in observational studies. This means that is not a domain task ontology which represents actors, models or processes. Potential users are researchers using CDs and biostatisticians especially the juniors. Its aim is not to achieve the most formal representation possible but rather to help in (i) the construction, (ii) the analysis and (iii) the understanding of CDs inference results needed for covariate selection. To do so, OntoBioStat acts as a minimal and sufficient graph data model able to answer eight competency questions (table 1). There are all routine questions that biostatisticians ask themselves when it comes to covariate selection. Similar questions can easily be found in the template published by Cochrane [29, 30] and from CD literature. The aim of covariate selection is to highlight confounders (q2 and q3), exclude subjects, for example, patients with a contraindication to a given treatment (q 1), use a proxy in place of a given confounder that is unmeasured (q 6), detect missing data that lead to bias the true causal effect (q 4), list existing remaining biases (q 8) and their direction (q5). The 7th question (q7) enables to understand how and why two variables are correlated. The answers correspond to inferences made with axioms and SWRL rules. Since answers are defined classes and object properties, queries were not used. In addition, all inferences can be explained because of transparent reasoning.

Table 1: OntoBioStat Competency Questions

q1) Are patients (observations) at risk of not being comparable at all?
q2) Which covariates may confound the causal path between exposure and outcome?
q3) Which interactions should be included?
q4) Which missing data may bias estimation of true effect?
q5) What is the direction of the bias caused by a confounder X?
q6) Are there any proxies of confounders that should be adjusted for?
q7) What type of relation exists between two variables?
q8) Is there any bias that cannot be corrected using variable selection?

Previous ontologies and DAGs

As exposed in the introduction, there are no ontologies for covariate selection in the causality framework. However, some classes or object properties inspired OntoBioStat design (i.e., some existing classes could be useful for covariate selection) as well as DAG literature and have been cited and incorporated in the subsequent description of OntoBioStat.

Enumeration of all important terms

A corpus of terms related to the information available for a given study (e.g., location or length of follow up) and CD vocabulary (e.g., shared ancestor, collider) was created. These terms were extracted manually from peer-reviewed journal articles including : (i) theoretical and pedagogical articles about CDs and DAGs for covariate selection, or bias identification and (ii) the most known reporting guide-lines from the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network site (https://www.equator-network.org/reporting-guidelines/) [31-35]. Each term was defined either by a short quote from the original article or by personal definition. Then at least one source was provided as a PubMed Identifier (PMID) and potential synonyms and/or antonyms were added. Concerning the terms extracted from the guidelines, only non-explicit terms were defined.

The corpus of 72 terms underwent a process of validation (semantic validation) through consensus among three experienced biostatisticians accustomed to the use of DAGs in observational studies (TPL, ELe, ELa). To be valid, a term had to be understandable by the expert biostatistician and its definition had to be clear enough (without ambiguities).

Ontologization

OntoBioStat is expressed using the Ontology Web Language (OWL) standard [36] and was built with Protégé 5.X [37] using: (i) the corpus and basic components of DAGs/CDs, (ii)

DAG features and tutorials for building DAGs, both found from the literature review, (iii) theoretical cases from CD representations (see example in Figure 1), (iv) the list of the competency questions (Table 1) and (v) existing ontologies. Concerning the three first diagrams represented in Figure 1 for mediator, collider and confounding variables respectively, definitions were provided in the introduction. The fourth represents a causal diagram enhanced with OntoBioStat classes which are Direct_confounder and Confounder_like. Both are confounders but in two different ways. The last diagram represents the so called M-bias because the diagram is shaped like an M. In case of collider selection, the generated bias can still be corrected by adjusting for one of the vertices (variable A or B).

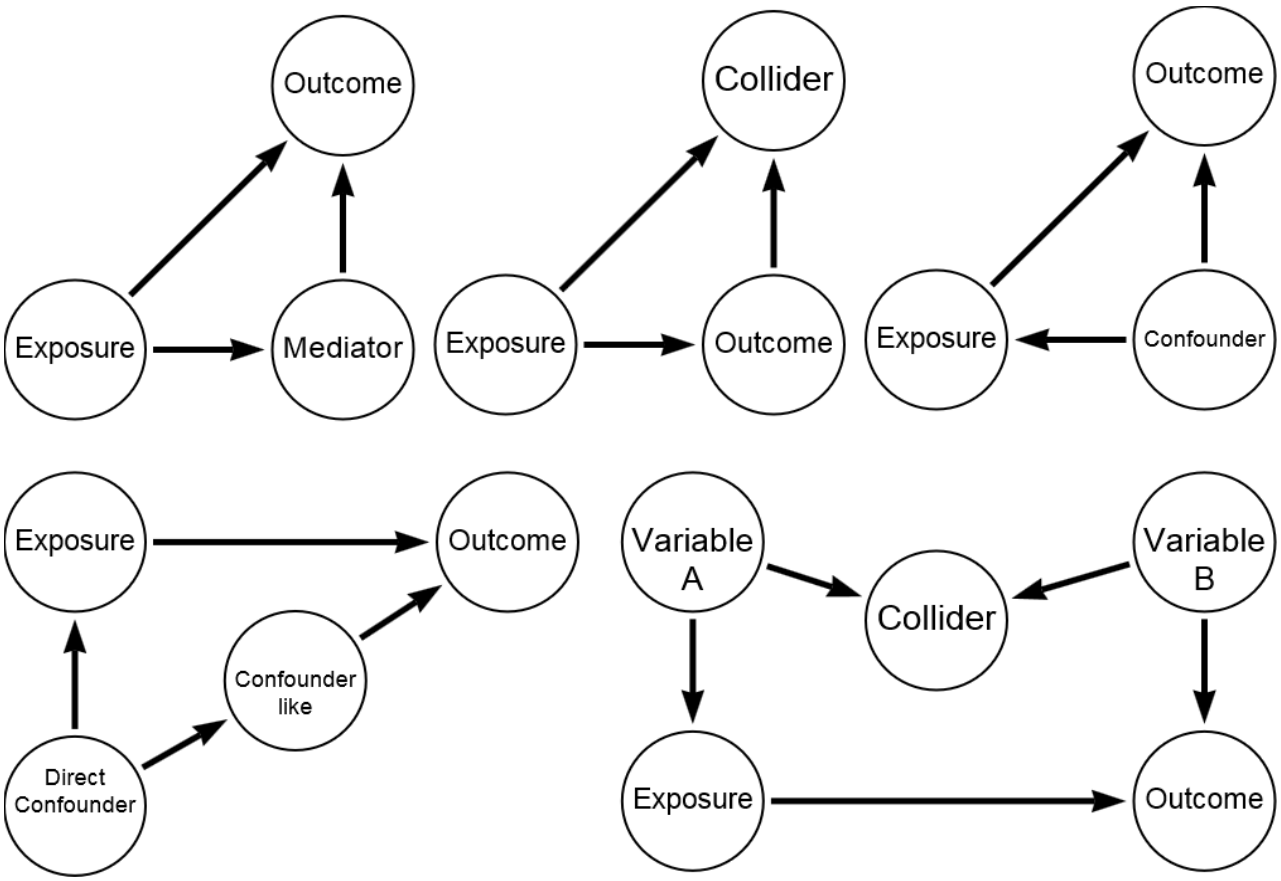


Figure 1: Theoretical cases of Causal Diagrams. From left to right and top to bottom: Mediator, Collider, Confounder, OntoBioStat classes and M-Bias.

Because OntoBioStat is an application ontology, it does not start with an upper-level ontology. For the while, OntoBioStat classes are not intended to be mapped with other ontology classes for the same reason.

Classes, *object properties*, and data properties created or reused correspond to theoretical medical knowledge (e.g., guidelines, medical textbooks), practical knowledge (e.g., protocol, machine used, staff, period, location, etc.) and data knowledge (e.g., type of variable, database, missing data, etc.). Classes and potential instances (i.e., represented knowledge) must be available in medical knowledge, study protocol, methodology or biostatistician report and dataset.

As explained in the introduction, edges and nodes are the primitive components of a CD. Hence, most of the instances of classes are variables (nodes) measured or not. These instances are linked with object properties corresponding to edges (e.g., causal relations). Datatype properties were preferred to classes when for a given instance there was no possible variation across study (e.g., the variable A instance of class Variable, has datatype properties hasType ‘quantitative’, i.e., A is a quantitative variable and A is always a quantitative variable).

After each modification of SWRL rules or axioms, the Pellet reasoner [38] has to be run on several theoretical instances sets and previous use cases in order to check for consistency. The lastest version of OntoBioStat is freely available online in the BioPortal ontology server at the following URL: https://bioportal.bioontology.org/ontologies/OBS.

OntoBioStat was presented following the **M**inimum **I**nformation for **R**eporting an **O**ntology **(**MIRO) guide-line [39]**.** MIRO is a guideline aiming at improving completeness and homogeneity in ontology reporting.

OntoBioStat

In a firs stage, OntoBioStat was built using entities that represents CD/DAG basic components in order to understand what it adds compared to classical construction framework and inferences. Then, the enriched context and the new inferences were introduced. OntoBioStat metrics are summarized in Table 2.

Table 2 : Metrics of the OntoBioStat ontology as reported in the Protégé tool

Axioms count	453	Annotation assertion	82
Declaration axioms	134	Logical axioms	237
Class count	57	Equivalent classes axioms	8
Object Properties count	35	SubClassOf axioms	55
Individuals count	30	Disjoint classes axioms	4
Data property count	11	SubObjectPropertyOf	27
Annotation property	1	InverseObjectProperties	3
		DisjointObjectProperties	1
		SymmetricObjectProperty	7
		AsymmetricObjectProperty	7
		IrreflexiveObjectProperty	7
		ObjectPropertyDomain	2
		ObjectPropertyRange	2
		ClassAssertion	60
		DataPropertyDomain	10
		DataPropertyRange	11
		SWRL rules	33

Basic DAG Components

In this part each term that represent basic component of a DAG was convert in an ontological term. Basic components of DAG/CD are exposure (**Exposure_stressor**), outcome (**Outcome**), other variables (**Covariates**), controlled (adjusted or adjustment) variable (**Path_Modifier**). **Path_Modifier** represents a broader category encompassing **Matched**, **Ponderated**, **Adjustment**, **Stratified**, and **Excluded**, which are various methods used to address and correct bias. **Path_Modifier** is a class needed for causal object properties reasoning and enables to explain results of a multivariate model, or subgroup analysis [40]. Unmeasured or unobserved variables are represented with the dataproperty is**Measured** with the value FALSE. Causal relations represented with CDs are represented in OntoBioStat: one asymetric irreflexive object property *isCauseof* (unidirectional edges), and two symmetric

irreflexive object properties *Share descendant* (non directional edge) and *share ancestor* (bidirectional edges). The list of object properties was extended in order to have a deeper understanding of correlations between variables and simpler SWRL rules (Fig 2).

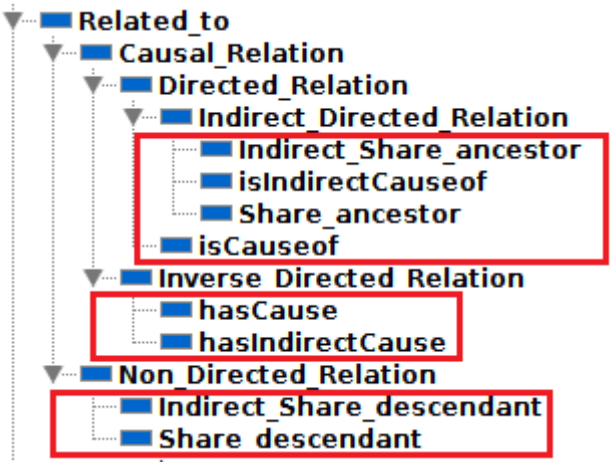


Figure 2: Causal hierarchy

Principal inferences on CDs or DAGs are based on the back door criterion [9] or similar algorithms that enable to select minimal or maximal confounding set of variables [10]. OntoBioStat highlights all possible confounders with a total of four subClasses of **Covariate** corresponding to the different lines of reasoning (i.e., axioms or rules) leading to the conclusion "confounder": **Collider confounder**, **Direct_Confounder**, **Indirect_Confounder**, **Confounder-like**. **Mediator** and **Collider** are inferred too.

In OntoBioStat, any variable that increases bias is considered as a 'confounder' (i.e, selection or measurement bias), hence other classes were created and detailed below. Furthermore, the same instance can be considered as a confounder for various reasons, which is why confounder classes are not disjoint classes.

Advanced DAG Components

In this part each term that represent advanced component of a DAG was convert in an ontological term. It includes: interaction, missing data mechanism and signed DAGs.

Including an interaction term between two variables is part of the variable selection process. Four different representations based on DAGs were found in the literature [41-44]. Inspired by the graphical Weinberg representation [42] two object properties (*Interact_with* and *isModifiedby*) and four subClasses of **Covariate** (**Pure_effect_modification**, **Interaction_Single**, **Synergistically_Antagonistically** and **Interaction_Confounder**) inferred with SWRL rules were created. **Interaction_Confounder** enable to answer to the 3rd question.

Handling missing data is a cornerstone of data analysis. Indeed depending of the mechanism that leads to missingness, estimates of true causal effect could be biased. Three mechanisms were defined by Rubin [45]: (i) missing completely at random (MCAR), (ii) missing at random (MAR), and (iii) missing not at random (MNAR). Covariate with missing data mechanism MCAR will be preferred instead of MNAR or MAR for covariate selection (corresponds to the 4th competency question: Which missing data may bias true effect?). Indeed, DAGs could be used to represent the three mechanisms [46] using two extra nodes: the reason for missing data and the true value of the missing data. Hence, in OntoBioStat, an adapted representation was created in order to infer MAR and MNAR subClasseof **Inferred_Missing_data_mechanism**. These are inferred based on causal relation *isCauseofNA* between the variable and an extra instance corresponding to the **Missing_value_reason**.

Sometimes, bias cannot be handled because of unmeasured confounders. However it is possible to estimate the direction of the bias thanks to signed DAGs [47-49]. Inspired by these representations, four object properties were created :

(i) *Increase*, *Decrease* are enable to answer the 5th question “What is the direction of the bias caused by a confounder X?”. For example, given a set of variables: **Exposure_stressor** (e),

Outcome (o) and **Covariate** (c). If c *Increase* e and *Decrease* o, the direction of the bias is *Decrease*.

(ii) *Contraindication* and *Absolute-indication* are sufficient causes closed to the monotonic effect definition. As defined in the Medical Subject Heading (MeSH) thesaurus, causes are sufficient when they initiate or produce an effect. These enable to infer **Unadjusted_Confounder** and answer the question "Should we exclude these patients?". For example, if we would like to compare thrombolysis and thrombectomy, time between stroke first symptoms and medical contact higher than 4 hours and 30 minutes *Contraindication* thrombolysis and isCauseof neurological prognosis. Hence, these patients with a medical contact higher than 4 hours and 30 minutes should be excluded.

Improve and Explicit context knowledge and inferences

In this part several elements are presented: classes concerning (i) context of a study (**Meta_Variable**), (ii) type of exposure and outcome (**Theoretical_Variable**), (iii) implicit variables considered as necessary cause of exposure or outcome (**Necessary_Variable**), dataproperties representing additional information about variable helping to assess data quality and finaly object properties that allow leveraging the uncertainty of knowledge.

Meta_Variable gathers classes corresponding to the knowledge about the **Variable** from a given study: (i) **Author point of view**, (ii) **Decision-makers (also a necessary variable)**, (iii) **Method of measurement (also a necessary variable)**, (iv) **Location**, (v) **Period**, (vi) **Collection_method** and (vii) **Missing value reasons**. These classes were inspired by GO-CAM [24], and the term extract from guide line reporting. They may be caused by other variables and are variables themselves. Hence, they may confound the true causal effect. For each one there is a corresponding object property (e.g., for the class **Period** there is an object property *isPeriodof*) grouped in *isMetavariableof* subObjectpropertyof *isCauseof*.

Theoretical_Variable groups all variables that can be considered as **Exposure_stressor** or **Outcome** under the five following sub-classes: (i) **Condition** (National Cancer Institute Thesaurus: **Disease, Disorder or Finding** [50]), (ii) **Environmental_Exposure** (from Human Health Exposure Analysis Resource (HHEAR) ontology [37]), (iii) **Health_Behavior** (groups 4 HHEAR classes: **Diet and Nutrition**; **Alcohol, Tobacco and Illicit Drug Use**; **Physical Activity and Fitness**; **Sleep Characteristic**), (iv) **Intervention_Effect** (HHEAR class: **Prescription Medication and Dietary Supplements**, Medical Action Ontology: **medical action** (https://github.com/monarch-initiative/MAxO)), (v) **Status** (groups 5 HHEAR classes **Demographic**, **Anthropometry**, **Dead**, **Socioeconomic Status**, **Medical History**). Using OntoBioStat, it becomes feasible to distinguish between the effect of a condition and the state of having a condition. HHEAR ontology gathers all exposome and supports implementation of exposure studies in existing ones.

Necessary_Variable groups all necessary or prerequisite variables for the realization of a given instance that could be **Exposure_stressor** or **Outcome**: **Exists**, **Available**, **Indicated**, **Prescribed**, **Delivered**, **Investigated**, **Measured**, and **Adhered**. **Necessary_Variable** classes are inspired by the ontology of causation made of state, event and process needed or allowing the realization of a given action [25]. According to the **Theoretical_Variable** stated by the user, various pre-included instances of **Necessary_Variable** and **Meta_Variable** will be linked through *isCauseof* object properties to the exposure and outcome using SWRL rules. These pre-include instances are mostly unmeasured covariates (Datatype property isMeasured FALSE), but they enable to infer if a covariate is a confounder through reasoning and prevent error building [52], see shorts examples below: (i) if we are interested in the effect of a given treatment on time before disease diagnosis. Treatment A side effect lead to consult more often physicians (decision maker) which leads to investigate more often symptoms and existing disease may be found

sooner (Figure 3); (ii) if we are interested in the effect of a given medical device on the disease development. With the medical device A it is easier to investigate the disease and hence investigation is realized more often. (Figure 4). In both cases building CDs without using OntoBioStat could lead to errors. Conventional algorithms may mistakenly identify variables as mediators, discouraging their inclusion, even when they actually contribute to bias. Two SWRL rules based on Necessary_Variable enable to infer **Mediator_Differential_Confounder** and **Differential_Bias_Unadjusted_Confounder**. Furthermore, the second example correspond to a bias that cannot be corrected (8th competency question). Finally, the last example enables to disentangle prescription from treatment effect as this is not the treatment effect that lead to co-prescription but only the prescription itself (Figure 5).

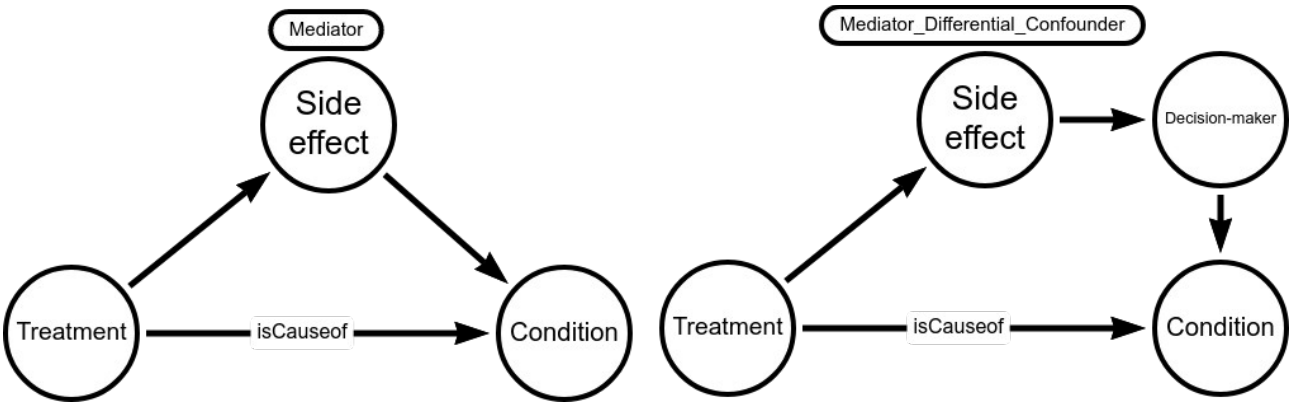


Figure 3: Standard causal diagram considers side effect as Mediator, Ontological causal diagram considers side effect as Mediator Differential Confounder

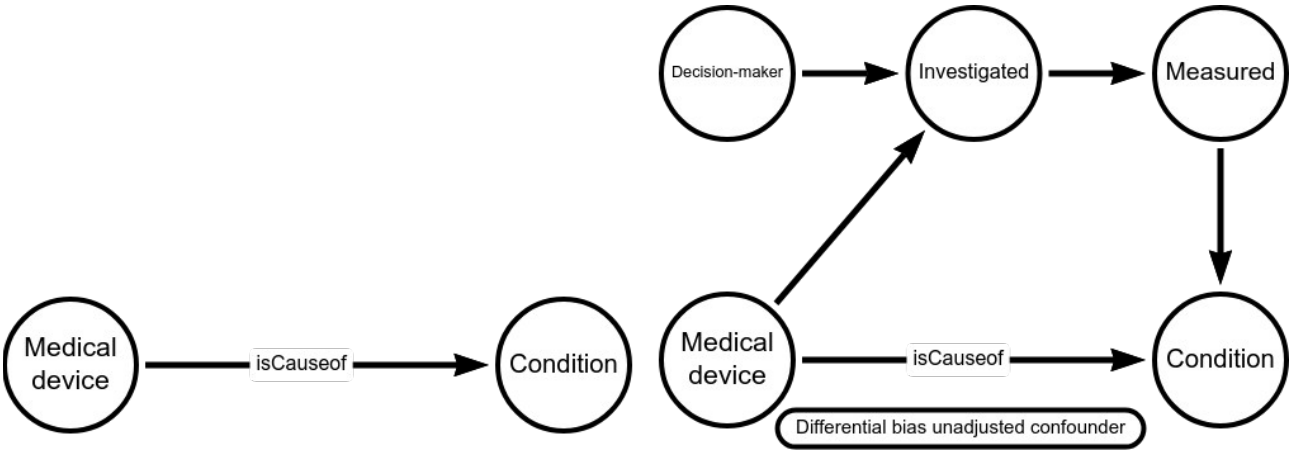


Figure 4: Standard causal diagram considers variable Investigated as invisible, then as a Mediator and finally with adequate inference as a Differential bias unadjusted confounder

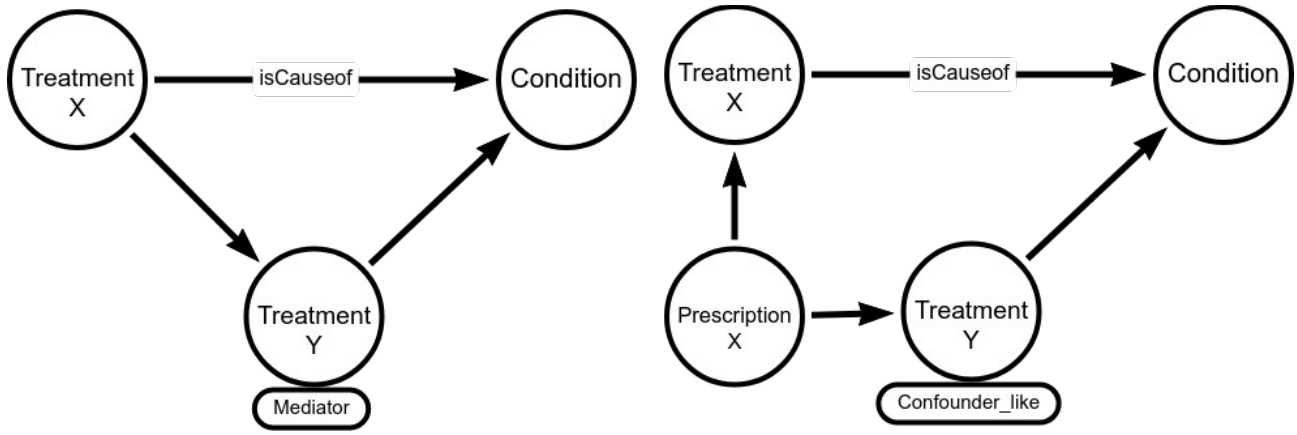


Figure 5: In a standard causal diagram representation variable Treatment Y is a Mediator, in an Ontological causal diagram is a Confounder-like (because of the node Prescription)

Various information about variables is available in the protocol or method part such as definition or source, summarized in the reporting guide lines. In order to represent all these additional information 10 dataproperties were created: (i) has**Definition;** (ii) has**Type** (dichotomous, ordinal, polychotomous, quantitative); (iii) has**Possible_value** represents possible values for a given variable ("Constant", "Bounded", "Complete","Incomplete"). "Constant" is a possible value for a variable: constants are often omitted because there are not proper variables (i.e., do not vary) but constant covariate is a **Path_Modifier**. Hence, if this is

a collider [53] or a mediator this constant covariate leads to bias the true effect. “Bounded” [54] is even more subtle because the covariate value varies but has not the full theoretical range. These covariates may be identified in the eligibility criteria section of a study (e.g., age >18). “Incomplete”: Incomplete variable is a variable that has been already transformed (e.g., age from 0 to 75 transformed in categorical age: [0-10][11-20][21-40] … etc) or simplified (Diabetes 1, 2 with insulin or not, no diabetes simplified in Diabetes yes or no). Inappropriate transformation or simplification may lead to loss of information; (iv) **Data_source** (e.g., “registry”, “electronic health report”, …): this one could become a metavariable with the increasing number of multicenter multisource studies; (v) **mayhaveMissing_value**: Covariable may be collected with or without missing data (e.g., by default simple imputation by 0 in healthcare claim database); (vi) has**Missing_value_amount** represents the percentage of missing value of a given instance. In case of complete case analysis (listwise analysis), included covariates with missing data in multivariate model results in losing observations when there is at least one missing value among the covariates included. Hence covariates with lower percentage of missing value will be preferred for covariate selection. There are four additional dataproperties about time knowledge that were not presented here. An example of knowledge enhancement with **Meta_Variable** and data properties about cerebral necrosis size is represented in the Figure 6.

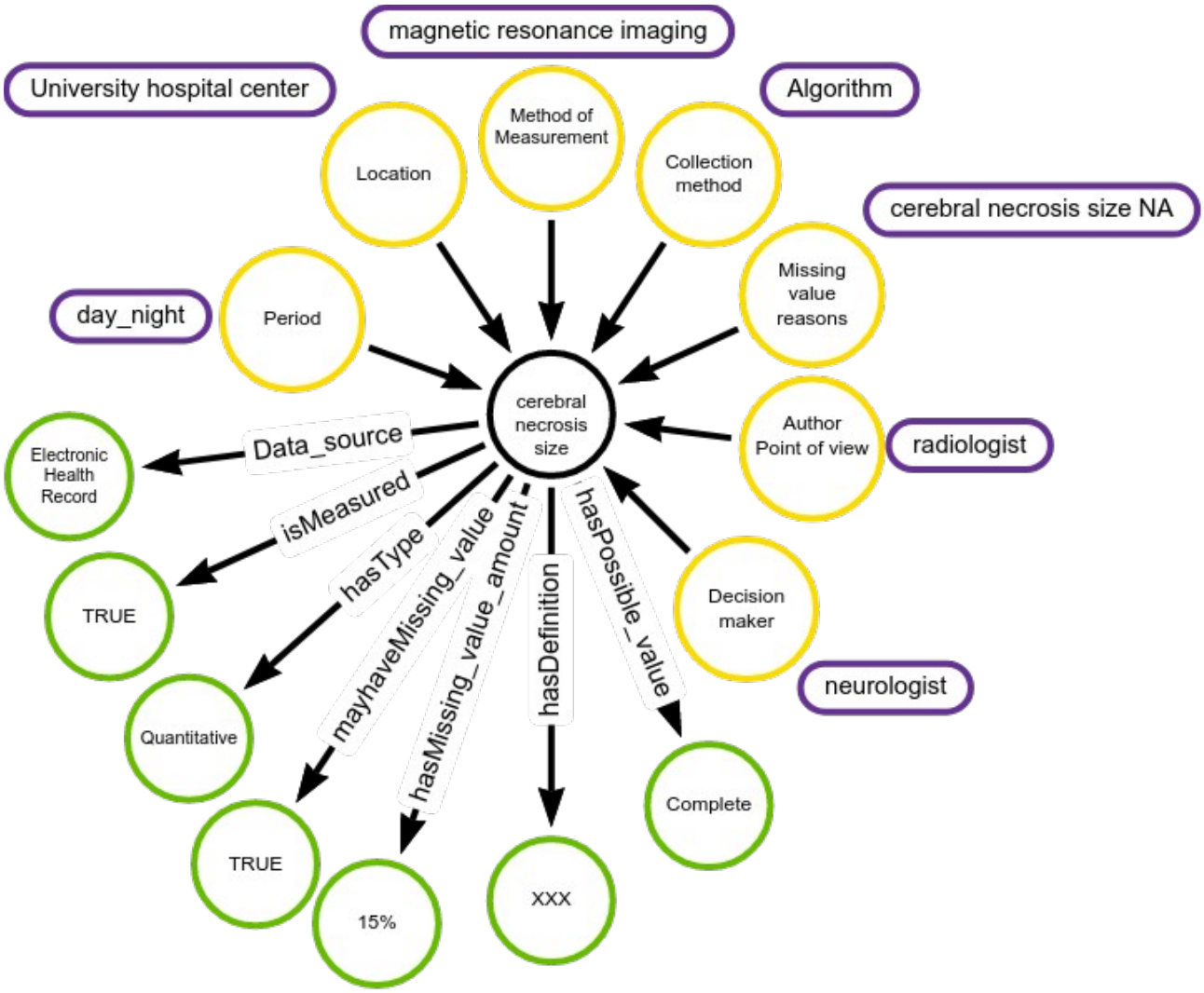


Figure 84: Variable cerebral necrosis size represented with all the Meta_Variable and Data properties

In order to leverage knowledge uncertainty, it is common practice to perform sensitivity analyses which correspond to the same model with some tweaks. Two causal relations (two *isCauseof subObjectproperties*) were added in order to represent this uncertainty: *Cause_Hypothetically* : pure speculation based on feeling or « clinical logic ». *Cause_Under_Validation* : causal relation found in articles but in need of more studies to confirm it.

In causal DAGs, the absence of edge between two variables means no causal relation. However, due to the open world assumption we made this statement explicit with the

NotCauseof object property useful for proxies' detection. Proxies, especially **Proxy_Confounder**, are variables that are ancestors or descendants of a confounder without being the cause of exposure and outcome. Proxies can be selected in place of an unmeasured confounder (6th competency question).

In some situations, the **Outcome** cause directly or indirectly the exposure. This is called **Reverse_Causality** and cannot be corrected by covariate selection. CD are by definition acyclic whereas OntoBioStat may represent cycles involving **Outcome** and **Exposure,** hence **Reverse_Causality**.

Workflow and use case

OntoBioStat is used thanks to Protégé an ontology editor that includes a reasoner with a user friendly interface. The different steps of ontological CD construction and analysis are summarized below (Figure 7). For more information about the user interface, see Appendix.

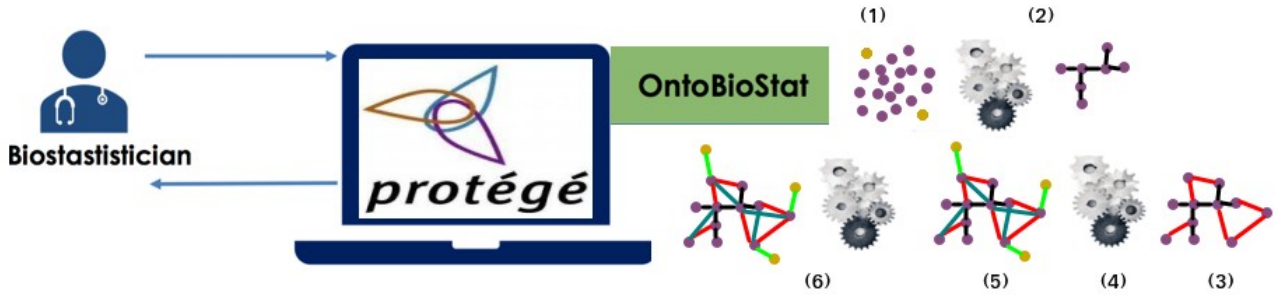


Figure 85: Figure summarizing the different steps when using OntoBioStat: (1) the user specifies the instances corresponding to the Exposure_stressor and the Outcome and their respective theoretical class, (2) the reasoner is activated and the automatic construction with the necessary variables occurs, (3) the user specifies the covariates and the causal relations, (4) the reasoner is activated, (5) the user refines the ontological causal diagram by removing and/or adding relations according to the inferences, (6) last inferences, reading and explanation of the results.

Study description:

In order to understand how we may translate the knowledge about a given study into a CD using OntoBioStat, the description of the translation (i.e., how to instantiate) will be presented together with the example study.

The research question was: “Does the surgical procedure resection and primary anastomosis (RPA) compared to RPA with protective stoma (**Exposure_stressor** and **Intervention_effect**) decrease the risk of severe surgical complications (**Outcome** and **Condition)** at 90 days? **Exposure_stressor** has**Definition** protective stoma consists of placing a protective stoma that must be removed later. This stoma protects against the effects of anastomosis leaks, which are likely to cause peritonitis, sepsis or even shock. **Outcome** Clavien-Dindo >= IIIb has**Definition** surgical revision or a life-threatening complication requiring intensive care within 90 days post operative including complications following stoma removal excluding immediate post-op complications”.

The study design is a retrospective multicenter (**Location**) observational study. Selected patients had a diverticular peritonitis (**Path_Modifier** diverticular peritonitis *hasCause* sex and age) with Hinchey classification between III and IV (has**Possible_Value** Bounded) and treated by RPA with or without protective stoma. Patients who underwent Hartman’s surgical procedure were excluded. Data collection was done manually by one physician per center based on medical report and observations (**Collection_method** ‘manual human’, has**Data_source** ‘EHR’).

Collected variables (**Covariate**) all measured (is**Measured** TRUE):

- MetaVariable **Period** night surgery.

- MetaVariable **Decision-maker** experience of the surgeon (senior, junior)/ specialty of the surgeon (generalist, specialist) impacts the choice between RPA with or without protective stoma (both has**Missing_value_amount** 10%).

- MetaVariable **Location** Location of the surgery (‘Center’) impacts choice between RPA with and without protective stoma and may have different post-operative protocol that impact **Outcome** (Center *Cause_Hypothetically* post-operative protocol*)*.

- ASA-Score (has**Missing_value_amount** 2%), an indicator of the patient's comorbidities as well as their condition before surgery (possible shock, etc) (*NotCauseof* **Outcome***)*. Value

depending on **Collection_Method** (there was no specific instruction whether it should be recalculated or not neither imputed or not) and **Author_point_of_view** (anesthesiologist). Missing data could be cause by the ASA-Score itself (*isCauseofNA* ASA-Score_NA).

- Body mass index (has**Missing_value_amount** 5%) indicator of nutritional health behaviour, sedentary lifestyle implicated in arterial disease and abdominal circumference (waist circumference). The thickness of the abdominal wall can make it difficult to wear a stomy and tissue healing depends on the health of the underlying arterial network.

- The variable diabetes should ideally have various values, including type 1 and type 2, with or without insulin treatment. However, since we only have access to a binary representation (yes or no) of diabetes, this variable has**Possible_Value Incomplete**. Furthermore, since the absence of diabetes is not always explicitly documented in medical records, the lack of any mention of diabetes has been treated as a genuine absence of the condition. This was addressed through native imputation with a value of 0 hence diabetes mayhave**Missing_value** FALSE.

- The variable immunosuppression may be due to treatment (e.g., corticosteroids, chemotherapy) or disease. Immunosuppression is associated (*Cause_Under_Validation*) with post operative abdominal collection and increases post-operative mortality (has**Possible_Value** Incomplete. mayhave**Missing_value** FALSE).

- Passage by an intensive care unit (ICU) at the end of the surgical procedure (has**Missing_value_amount** 10%). This ICU stay is not caused by one type of surgery rather than another and is not responsible for a Clavien-Dindo ≥ IIIb (*NotCauseof* **Outcome**).

- Age could eventually lead to prefer a procedure over another and aging implies more frequent post operative complication.

- Sex is not related to procedure choice neither the **Outcome**.

Unmeasured variables (isMeasured FALSE) :

- The initial hemodynamic instability or organ failure was not measured. However, in our subgroup it should be constant (has**PossibleValue** Constant) as hemodynamic instability is a contraindication to RPA (hemodynamic instability *Contraindication* RPA). In a small percentage of cases, depending on the degree of instability, an RPA would have been performed instead of a Hartman. The object property signed *Contraindication* will therefore not be selected in our use case.

The ASA score from level 4 onwards could correspond to a failure state (failure *isCauseof* ASA). Ending in ICU after the surgical procedure is a consequence of an existing or latent failure state at the beginning of the surgery (failure or risk of failure *isCauseof* ICU stay).

All the following unmeasured variables impact the **Outcome** and the first two variables could impact the **Exposure_stressor**: abdominal wall thickness, arteritic disease, wound healing, post-operative protocol (antibiotic therapy), rehabilitation.

Concerning interactions, none were described in this study.

Concerning the direction of the bias, signed object properties were not used, because of the absence of contraindication or absolute indication. Moreover, *Increase* and *Decrease* properties were not used because covariates causing the outcome also increased the outcome, and all covariates causing the exposure (except for abdominal_wall_thick) also decrease the exposure. Causal path do not contain too much nodes hence, the direction of bias caused by a confounder is easy to anticipate without inferences (Question n°5).

The different steps of reasoning (summarized in Figure 7) will be presented in the results. Only relevant classes, object properties, and data properties were selected for this use case.

Results

1) In the same way that research typically begins with the formulation of a research question, the specification of the exposure of interest, and the identification of the outcome,

OntoBioStat requires us to clearly define which instances represent the **Outcome** and the **Exposure_stressor**. Each of these instances must also be classified under a Theoretical Variable, in this use case **Condition** and **Intervention_Effect**, respectively.

2) Activating the reasoner for the first time helps us automatically build an initial diagram with the necessary variables. These variables are generic instances, meaning they can be used in various scenarios. Depending on the theoretical classification of the Outcome and the Exposure_stressor, OntoBioStat uses a subset of the necessary variables to create an initial graph. In this graph, these necessary variables are connected either to the Outcome or to the Exposure using causal relation. The necessary variables were not useful in our use case because, whether it is the exposure or the outcome, these are two variables that correspond to sporadic events (surgical treatment is not a long-term treatment delivered by a pharmacist), and the outcome is clinically noisy hence there is no differential lag bias (i.e., no major delay between the appearance of the disorder and its diagnosis). There will therefore be small risk or no risk of follow-up or classification bias for either the exposure or the outcome. This means that the necessary variables will not be causally related to any other variable than exposure and outcome. From a formal point of view, variable impacting treatment choice should cause the necessary variable Indicated, so it would be legitimate to want to use the necessary variables.

3) Then, instances corresponding to the **Covariates** were specified. Following that, context were defined using the **Meta_Variables**. Causal relationships between these different instances were defined. Additional information were provided through data properties. A total of 26 instances (15 **Covariates**, 11 **Metavariable** of which five **Missing_Value_Reasons**), 50 object properties and 37 data properties were asserted using Protégé (Appendix for the initial graph, only instances and object properties).

4) Reasoner was activated allowing to identify potentially useful proxies.:

A total of 12 Confounders were inferred Axioms and SWRL rules enabled to answer to the competency question (see Table 4).

In order to understand inference related to proxy, we relied on the explaining inference feature. An example of explanation is provided in the Figure 8, in which the instance ICU-stay is classified as a **Proxy_Confounder** thanks to several assertion and SWRL rule (line 5). The following proxies were detected: ASA-score (for arteritic condition and initial hemodynamic instability); ICU-stay (for initial hemodynamic instability); BMI (for abdominal thickness and arteritic condition). ASA-Score is close to arteritic condition (isCauseof) contrary to BMI (Share_ancestor), hence ASA-Score was preferred instead of BMI. However, as BMI is a good proxy for abdominal thickness, it was used as proxy too. ASA-score had few missing data, and depends of author point of view and data collection. ICU-stay had missing data depending of the Center where the data were collected (MAR: Missing At Random). BMI had missing data that seemed to be missing completely at random.

5-6) Unmeasured confounder instances (i.e., arteritic condition, initial hemodynamic instability and abdominal thickness) were replaced by proxies confounder and the reasoner was activated. Results were identical to thus shown in the Table 4 except for Proxy_Confounder becoming Indirect_Confounder and unmeasured confounders that were erased. Finally, depending on the drawn graph, there were no bias that cannot be corrected using variable selection (question 8), and there were no patients who needed to be excluded (question 1). As no object properties related to interaction were included, there were no inferences about it (question 3). As state, in the method part, signed object properties were not used however, table 4 was filled with expected results (question 5).

Sensitivity analysis with alternative set of confounders: Considering the retrieved confounders; Immunosuppression was Incomplete (low quality variable) and the causal relation with the **Outcome** is under validation. Surgeon information and center location were

considered as confounder because of the hypothetical causal link between post operative protocol and the Outcome. These variables are considered as confounders because of these uncertain causal links, so we suppressed the hypothetical and under validation causal link and reran the reasoner in order to obtained a different set of confounders.

Table 4 : Competency questions, answers and explanations

Competency Questions	Use Case answers	Classes
1-Are patients (observations) at risk of not being comparable?	No	**Unadjusted_Confounder**
2-Which covariates may confound the causal path between exposure and outcome?	Age, sex,arteritic condition, ASA-Score, wound healing, initial hemodynamic instability, post-operative protocol, surgeon experience, surgeon specialty, abdominal wall thick, Center, immunosuppression.	**Confounder_like**, **Direct_Confounder**, **Indirect_Confounder**, **Mediator_Confounder**, **Interaction_Confounder**, **Proxy_Confounder**, **Unadjusted_Confounder**, **Collider_Confounder**.
3-Which interaction should be included?	Not concerned	Interaction_Confounder
4-Which missing data may biased true effect?	ASA-Score is MNAR, ICU stay, surgeon experience and specialty are MAR	**MNAR**, **MAR** and **MCAR**
5-What is the direction of the bias caused by a confounder X?	Confounding tend to bias true causal effect downward except for abdominal_wall_thick. Hence the new surgical procedure may appear protective of Clavien-Dindo >= IIIb.	Product of signed object properties (*Increase*/*Decrease*) between: (i) Covariates that may confound the causal path and (ii) the **Outcome** and the **Exposure_stressor**. e.g: the **Covariate** 'cancer' increases the **Outcome** 'Clavien-Dindo >= IIIb' and decrease the **Exposure_stressor** 'surgical_procedure'. Hence, 'cancer' biased the true causal effect downward.
6-Are there any proxy of confounders that should be adjusted for?	ASA-Score, ICU_stay, BMI	Proxy_Confounder
7-What type of relation exists between two variables?	1465 object properties were inferred to summarize different ways of defining relations.	All Related_to sub-properties. e.g: (i) Hinchey *isIndirectCauseof* RPA, (ii) ASA_score *Share_ancestor* with ICU stay, (iii) wound healing *Related_to* ASA-Score
8- Is there any bias that cannot be corrected using variable selection?	No	**Differential_Bias_Unadjusted_Confounder, Reverse_Causality**

OntoBioStat enables the representation of all available knowledge on the subject, whether it's medical knowledge, on-ground insights, study-related information, or data quality (including

missing data). This comprehensive representation is made possible through the framework provided by OntoBioStat within the Protégé tool.

Figure 86: Inference explanations from Protégé about ASA Score instance, inferred class Proxy_Confounder

4. Discussion

This proposed method aimed firstly to gather knowledge needed by a biostatistician in order to optimize covariates selection for causal inference in observational studies. This knowledge was ontologized in OntoBioStat, an application ontology capable of reasoning and supporting causal diagram construction for covariate selection. This use case has illustrated how to translate knowledge into ontological CD and answered competency questions. While the success of rich knowledge representation for variable selection in the use case is evident, it has not been demonstrated that OntoBioStat is inherently "better" than the absence of tools.

OntoBioStat succeeded in the implementation of the classical representation of Mediator, Collider, Confounder, Interaction and the various relations existing in DAG (e.g., bidirectional, non-directional, unidirectional). Contrary to the DAG representation,

OntoBioStat framework offers explicit knowledge representation with Meta and Necessary Variable classes, object property allowing to appreciate knowledge uncertainty and the dataproperties, which allow to gauge the quality of the variables in order to carry out sensitivity analyses with several different sets of variables. These classes and object properties were inspired by various ontologies and reporting-guidelines.

Concerning the inferences:

There are specific inferences based on sufficient (i.e., skill issue 1) and necessary causes (confounding variables inferred through necessary variables) that CDs cannot therefore resolve. Reverse causality cannot be represented either because CDs are acyclic.

Proxy inference requires knowledge of the measured or unmeasured status of a given variable and the absence of any causal links between the proxy variable and the outcome or exposure. It would be possible to infer proxy from a CD via a graph operation, but the *NotCauseof* representation of OntoBioStat help to validate the identification of proxies by making explicit the absence of causal links between proxy and the outcome or the exposure. In the use case, three proxies were highlighted and used in the model.

Considering that the biases caused by variables with missing data have to be highlighted to correctly select variables, OntoBioStat mixes causality for variable selection and causality for the understanding of missing data generation. In this use case, missing data for some variables could depend of the center, which is a current situation and not a necessary useful information considering the low amount of missing data, and the absence of other missing data reason that could bias the true causal effect. Considering the ASA-Score with 2% of missing data and MNAR missing data mechanism, significant bias should not be observed.

The Pellet reasoner provides a consistent ontology. SWRL rules and axioms enable to infer various classes and properties, and inferences are explained easily with the Protégé tool. In the use case, only some inferences were explained in order to verify reasoning however for a non familiar user this feature could be translated in natural language.

Concerning construction, dagitty [12] is one of the most advanced web based user friendly tools available for DAG building and analysis. However, visual approaches for CD building become quickly inefficient with the growing number of variables and causal relations. OntoBioStat construction and reasoning framework is a more complete approach than dagitty, which relies on basic component of DAGs only. In the use case, compared to a DAG, causal relation uncertainty, data quality and missing data were integrated to the knowledge representation. However, with 28 instances, manual entries of all statement through Protégé editor were clearly a limit and could be even more time consuming with more variables, that is why there is a need to enhance the usability with default settings such as all variable are considered is**Measured** TRUE and has**PossibleValue** Complete. Furthermore, we could rely on existing structured knowledge and extract this knowledge such as in [55] that query knowledge graph in order to identify confounders, mediators and colliders. In the use case, after all the specifications were made, some nodes were irrelevant such as night surgery. Irrelevant nodes could be erased from the ontological causal diagram applying a reduction method using SPARQL queries and/or graph operation. But ontological causal diagram could be still hard to read and deleting knowledge already specified should be considered as a real loss. For this use case, reasoning was quick enough (<30 s) to iteratively refine the ontogical causal diagram. Finally, we did not design a study where we compared standard CDs drawn by DAG experts with the ontology-based CD, but this should constitute an upcoming study.

Concerning the variable selection process OntoBioStat only highlights potential confounder and cannot answer the most important question that is: ‘What is the sufficient set of confounders? (that should correct all bias)’ with only natives SWRL rules or axioms. Solving this question requires path analysis with additional non-native functions developed in Java language. Dagitty and dagR [56] are both R packages that enable the selection of a sufficient or minimal set of confounders. Hence, in the study related to the use case Dagitty was used as a complementary tool. Furthermore, even if OntoBioStat could a provide minimal

set of confounders, sensitivity analysis with alternative variable selection should be performed by the OntoBioStat user without any mention of confidence degree from the ontology.

Concerning the corpus constitution, only some of the EQUATOR reporting guidelines about observational studies and randomized clinical trial were screened for term extraction. Indeed, guide-lines are redundant and only the most known were read. Furthermore, those about clinical trials brought details about ‘intervention’ which were used and could be used in further development. These latter enable to extend the application of OntoBioStat to randomized controlled trials and predictive studies. An up-to-date systematic review of CD in medical research and terms extraction were done only by one reader (TPL) but the corpus was validated by two others DAG/CD users (ELev and ELan). Articles from non-epidemiological journals were excluded because we considered information de facto redundant with journal specialized in epidemiology or biostatistics. Articles about DAG could be found out of PubMed bibliographic database (e.g., Embase). However, terms concerning DAG features in epidemiology should be nearly exhaustive in PubMed.

In order to criticize OntoBioStat we relied on the OOPS tool [57]. OOPS scans an ontology to detect various types of pitfalls and problems, such as modeling errors, logical inconsistencies, and design issues. Here is, reasons why pitfalls were detected: None of the entities had a label because the class name is already explicit. Although **Path_Modifier** and **Inferred_Missing_data_mechanism** are disconnected from the others entities they are inferred based on statements from others entities and SWRL rules. Some of the inverse asymetrics object properties are not declared because there is no need for the application. There is no specific naming convention except being understandable, we also added definition and example for each class. Domain and range are not specified for some of the object properties because all instances created in OntoBioStat could be domain or range.

5. Conclusion and some perspectives

Overall, OntoBioStat's sophisticated design showcases its holistic approach to incorporating the diverse facets of variable selection process into a unified framework. OntoBioStat addresses competency questions, yet there remains a need to enhance its usability and expand its range of competencies. R is one of the most used software among biostatisticians and includes packages for minimal set selection. Hence, combining the use of OntoBioStat and the R package dagitty could be fruitful, as it may allow: (i) performing path analysis using existing algorithms, (ii) solving bigger and complex graphs, and (iii) enhancing user experience .

Funding sources:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Contributors:

E.Le and E.La, DAGs users, validated and corrected the corpus used to build the ontology, and criticized the ontologisation process (conception and interpretation).

J.P made substantial contribution to the design of the ontological causal diagram from the use case and to the inference interpretation.

J.G, SJ.D, L.S and J.B made substantial contributions to the conception and design of the study and results interpretation.

T.PL designed the study, interpreted the results, and drafted the first version of the article.

All authors have made substantial contributions to all of the following: (1) revising the article critically for important intellectual content, (2) final approval of the version to be submitted.

Declaration of Generative AI and AI-assisted technologies in the writing process:

During the preparation of this work the authors used Chatgpt in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1- Hall NS. R. A. Fisher and his advocacy of randomization. J Hist Biol. 2007 Jun 25;40(2):295–325.

2- Robins JM, Morgenstern H. The foundations of confounding in epidemiology. Computers & Mathematics with Applications. 1987;14(9–12):869–916.

3- Joiner BL. Lurking Variables: Some Examples. The American Statistician. 1981 Nov;35(4):227.

4- Rutter M. Epidemiological methods to tackle causal questions. International Journal of Epidemiology. 2009 Feb 1;38(1):3–6.

5- Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet. 2002 Jan;359(9302):248–52

6- Howards PP, Schisterman EF, Poole C, Kaufman JS, Weinberg CR. 'Toward a clearer definition of confounding' revisited with directed acyclic graphs. Am J Epidemiol. 2012 Sep 15;176(6):506–11.7.

7- VanderWeele TJ. Mediation and mechanism. Eur J Epidemiol. 2009;24(5):217–24.

8- Westreich D. Berkson's bias, selection bias, and missing data. Epidemiology. 2012 Jan;23(1):159–64.

9- Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan;10(1):37–48.

10- Pearl J. Causality: models, reasoning, and inference. Cambridge, U.K. ; New York:

Cambridge University Press; 2000. 384 p

11- VanderWeele TJ, Shpitser I. A New Criterion for Confounder Selection. Biometrics. 2011 Dec;67(4):1406–13.

12- Ankan A, Wortel IMN, Textor J. Testing Graphical Causal Models Using the R Package "dagitty". Current Protocols. 2021 Feb.

13- Pressat-Laffouilhère T, Jouffroy R, Leguillou A, Kerdelhue G, Benichou J, Gillibert A. Variable selection methods were poorly reported but rarely misused in major medical journals: Literature review. Journal of Clinical Epidemiology. 2021 Nov;139:12–9.

14- Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. International Journal of Epidemiology. 2021 May 17;50(2):620–32.

15- Barnard-Mayers R, Childs E, Corlin L, Caniglia EC, Fox MP, Donnelly JP, et al. Assessing knowledge, attitudes, and practices towards causal directed acyclic graphs: a qualitative research project. Eur J Epidemiol. 2021 Jul;36(7):659–67.

16- Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. Journal of Clinical Epidemiology. 2022 Feb;142:264–7.

17- Ferguson KD, McCann M, Katikireddi SV, Thomson H, Green MJ, Smith DJ, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. International Journal of Epidemiology. 2020 Feb 1;49(1):322–9.

18- Kleinberg S, Hripcsak G. A review of causal inference for biomedical informatics. J Biomed Inform. 2011 Dec;44(6):1102-12

19- Ontologies Guarino N., Oberle D., Staab S. (2009) What Is an Ontology?. In: Staab S., Studer R. (eds) Handbook on Ontologies. International Handbooks on Information Systems. Springer, Berlin, Heidelberg.

20- Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of

Clinical Research (OCRe): An informatics foundation for the science of clinical research. Journal of Biomedical Informatics. 2014 Dec;52:78–91.

21- Bandrowski A, et al. The Ontology for Biomedical Investigations. PLoS ONE. 2016 Apr 29;11(4):e0154556.

22- Zheng J, Harris MR, Masci AM, Lin Y, Hero A, Smith B, et al. The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis. J Biomed Semant. 2016 Dec;7(1):53.

23- 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.

24- Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, Balhoff JP, Albou LP, Good B, Gaudet P, Lewis SE, Mungall CJ. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet. 2019 Oct;51(10):1429-1433.

25- Galton A. States, processes and events, and the ontology of causal relations. Frontiers in Artificial Intelligence and Applications. 2012 Jan;239:279–92.

26- Behnaz A, Bandara M, Rabhi FA, Peat M. A Statistical Learning Ontology for Managing Analytics Knowledge. In: Mehandjiev N, Saadouni B, editors. Enterprise Applications, Markets and Services in the Finance Industry p. 180–94. (Lecture Notes in Business Information Processing; vol. 345).

27- Nayak A, Božić B, Longo L. An Ontological Approach for Recommending a Feature Selection Algorithm. In: Di Noia T, Ko IY, Schedl M, Ardito C, editors. Web Engineering. Cham: Springer International Publishing; 2022. p. 300–14. (Lecture Notes in Computer Science; vol. 13362).

28- Noy, N. F. & McGuinness, D. L. (2001), 'Ontology Development 101: A Guide to Creating Your First Ontology' http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html

29- Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;i4919.

30- Stone JC, Glass K, Clark J, Ritskes-Hoitinga M, Munn Z, Tugwell P, et al. The MethodologicAl STandards for Epidemiological Research (MASTER) scale demonstrated a unified framework for bias assessment. Journal of Clinical Epidemiology. 2021 Jun;134:52–64.

31- Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007 Oct 20;335(7624):806–8.

32- Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010 Mar 23;340(mar23 1):c332–c332.

33- Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 2013 Jan 8;346:e7586.

34- Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ. 2018 Nov 14;k3532

35- Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014 Mar 7;348(mar07 3):g1687–g1687.

36- Bock, A. Fokoue, P. Haase, R. Hoekstra, I. Horrocks, A. Ruttenberg, U. Sattler, M. Smith, OWL 2 Web Ontology Language, W3C recommendation

37- Protégé: Musen, M.A. The Protege project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence,

1(4), 2015 Jun.

38- Pellet: Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y. Pellet: A practical OWL-DL reasoner. Web Semantics. 2007 Jun;5(2):51-53.

39- Matentzoglu N, Malone J, Mungall C, Stevens R. MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semant. 2018 Dec;9(1):6.

40- Pressat Laffouilhère, T. et al. (2022). Ontological Representation of Causal Relations for a Deep Understanding of Associations Between Variables in Epidemiology. In: Michalowski, M., Abidi, S.S.R., Abidi, S. (eds) Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science(), vol 13263. Springer, Cham.

41- VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology. 2007;18:561–8.

42- Weinberg CR. Can DAGs clarify effect modification? Epidemiology. 2007;18:569–72.

43- Nilsson A, Bonander C, Strömberg U, Björk J. A directed acyclic graph for interactions. International Journal of Epidemiology. 2021 May 17;50(2):613–9.

44- Lopez PM, Subramanian SV, Schooling CM. Effect measure modification conceptualized using selection diagrams as mediation by mechanisms of varying population-level relevance. Journal of Clinical Epidemiology. 2019 Sep;113:123–8.

45- Rubin D, Inference and missing data. Biometrika. 1976 Dec 1;63(3):581–592

46- Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. Stat Methods Med Res. 2012 Jun;21(3):243–56.

47- VanderWeele TJ, Hernán MA. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. Am J Epidemiol. 2012 Jun 15;175(12):1303-10.

48- VanderWeele TJ, Robins JM. Signed directed acyclic graphs for causal inference. J R Stat Soc Series B Stat Methodol. 2010 Jan 1;72(1):111-127.

49- VanderWeele TJ, Hernán MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured confounding bias. Epidemiology. 2008 Sep;19(5):720-8.

50- de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, Quan SL, Safran T, Thomas N, Whiteman L. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. 2009 Jun;42(3):530-9.

51- Viet SM, Falman JC, Merrill LS, Faustman EM, Savitz DA, Mervish N, Barr DB, Peterson LA, Wright R, Balshaw D, O'Brien B. Human Health Exposure Analysis Resource (HHEAR): A model for incorporating the exposome into health studies. Int J Hyg Environ Health. 2021 Jun;235:113768.

52- Pressat Laffouilhère T, Grosjean J, Bénichou J, Darmoni SJ, Soualmia LF. OntoBioStat: Supporting Causal Diagram Design and Analysis. Stud Health Technol Inform. 2022 May 25;294:302-306.

53- BERKSON J. Limitations of the application of fourfold table analysis to hospital data. Biometrics. 1946 Jun;2(3):47-53.

54- Panov P, Soldatova LN, Džeroski S. Generic ontology of datatypes. Information Sciences. 2016 Feb;329:900–20.

55- Malec SA, Taneja SB, Albert SM, Shaaban CE, Karim HT, Levine AS, et al. Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: a use case studying depression as a risk factor for Alzheimer’s disease. Systems Biology; 2022 Jul

56- Breitling LP. dagR: a suite of R functions for directed acyclic graphs. Epidemiology. 2010 Jul;21(4):586-7.

57- Poveda-Villalón M, Gómez-Pérez A, Suárez-Figueroa MC. OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. International Journal on Semantic Web and Information Systems. 2014 Apr 1;10(2):7–34.

Les figures ont été réalisées avec https://arrows.app.

Modèle ontologique formel, un appui à la sélection des variables pour la construction des modèles multivariés

Résumé

Répondre à une question de recherche causale dans un contexte d'étude observationnelle nécessite de sélectionner des variables de confusion. Leur intégration dans un modèle multivarié en tant que co-variables permet de diminuer le biais dans l'estimation de l'effet causal de l'exposition sur le critère de jugement. Leur identification est réalisée grâce à des diagrammes causaux (DCs) ou des graphes orientés acycliques. Ces représentations, composées de nœuds et d'arcs orientés, permettent d'éviter la sélection de variables qui augmenteraient le biais, comme les variables de médiation et de collision. Les méthodes existantes de construction de DCs manquent cependant de systématisme et leur représentation de formalisme, d'expressivité et de complétude. Afin de proposer un cadre de construction formel et complet capable de représenter toutes les informations nécessaires à la sélection des variables sur un DC enrichi, d'analyser ce DC et surtout d'expliquer les résultats de cette analyse, nous avons proposé d'utiliser un modèle ontologique enrichi de règles d'inférences. Un modèle ontologique permet notamment de représenter les connaissances sous la forme de graphe expressif et formel_composé de classes et de relations similaires aux nœuds et arcs des DCs. Nous avons développé l'ontologie OntoBioStat (OBS) à partir d'une liste de questions de compétence liée à la sélection des variables et de l'analyse de la littérature scientifique relative aux DCs et aux ontologies. Le cadre de construction d'OBS est plus riche que celui d'un DC, intégrant des éléments implicites tels que les causes nécessaires, contextuels d'une étude, sur l'incertitude de la connaissance et sur la qualité du jeu de données correspondant. Afin d'évaluer l'apport d'OBS, nous l'avons utilisée pour représenter les variables d'une étude observationnelle publiée et avons confronté ses conclusions à celle d'un DC. OBS a permis d'identifier de nouvelles variables de confusion grâce au cadre de construction différent des DCs et aux axiomes et règles d'inférence. OBS a également été utilisée pour représenter une étude rétrospective en cours d'analyse : le modèle a permis d'expliquer dans un premier temps les corrélations statistiques retrouvées entre les variables de l'étude puis de mettre en évidence les potentielles variables de confusion et leurs éventuels substituts ("proxys"). Les informations sur la qualité des données et l'incertitude des relations causales ont quant à elles facilité la proposition des analyses de sensibilité, augmentant la robustesse de la conclusion de l'étude. Enfin, les inférences ont été expliquées grâce aux capacités de raisonnement offertes par le formalisme de représentation d'OBS. À terme OBS sera intégrée dans des outils d'analyse statistique afin de bénéficier des bibliothèques existantes pour la sélection des variables et de permettre son utilisation par les épidémiologistes et les biostatisticiens.

Mots clé : Ontologie, Intelligence Artificielle, Diagramme Causaux, Graphe Orienté Acyclique, Causalité, Médecine, Sélection des variables, Facteur de confusion

Responding to a causal research question in the context of observational studies requires the selection of confounding variables. Integrating them into a multivariate model as co-variables helps reduce bias in estimating the true causal effect of exposure on the outcome. Identification is achieved through causal diagrams (CDs) or directed acyclic graphs (DAGs). These representations, composed of nodes and directed arcs, prevent the selection of variables that would introduce bias, such as mediating and colliding variables. However, existing methods for constructing CDs lack systematic approaches and exhibit limitations in terms of formalism, expressiveness, and completeness. To offer a formal and comprehensive framework capable of representing all necessary information for variable selection on an enriched CD, analyzing this CD, and, most importantly, explaining the analysis results, we propose utilizing an ontological model enriched with inference rules. An ontological model allows for representing knowledge in the form of an expressive and formal graph consisting of classes and relations similar to the nodes and arcs of Cds. We developed the OntoBioStat (OBS) ontology based on a list of competency questions about variable selection and an analysis of scientific literature on CDs and ontologies. The construction framework of OBS is richer than that of a CD, incorporating implicit elements like necessary causes, study context, uncertainty in knowledge, and data quality. To evaluate the contribution of OBS, we used it to represent variables from a published observational study and compared its conclusions with those of a CD. OBS identified new confounding variables due to its different construction framework and the axioms and inference rules. OBS was also used to represent an ongoing retrospective study analysis. The model explained statistical correlations found between study variables and highlighted potential confounding variables and their possible substitutes (proxies). Information on data quality and causal relation uncertainty facilitated proposing sensitivity analyses, enhancing the study's conclusion robustness. Finally, inferences were explained through the reasoning capabilities provided by OBS's formal representation. Ultimately, OBS will be integrated into statistical analysis tools to leverage existing libraries for variable selection, making it accessible to epidemiologists and biostatisticians.

Key words : Ontology, Artificial Intelligence, Causal Diagrams, Directed Acyclic Graph, Causality, Medicine, Variable Selection, Confounding Factor