Modèle ontologique formel pour la sélection des variables dans les modèles multivariés pour les études observationnelles en recherche biomédicale

Pressat-Laffouilhère, T., Grosjean, J., Darmoni, S., Benichou, J.

Les études observationnelles

Ces études apportent un éclairage sur des questions de recherches en vie réelle auxquelles les études randomisées ne peuvent répondre soient à cause d'une population non représentative ou bien encore parce qu'il ne serait pas éthique de randomiser l'assignation de certains traitements. Les études observationnelles, contrairement aux études randomisées, doivent faire face à des biais de confusion qui peuvent être majeurs remettant en doute la véracité des résultats [1]. En effet, en vie réelle, les examens et les traitements ne sont pas prescrits au hasard. Ils dépendent des caractéristiques des patients et aussi du médecin.

Les questions de recherche

La problématique de la sélection des variables dans les modèles statistiques repose en premier lieu sur la question de recherche. Le but est-il de prédire avec un ensemble de variables (par exemple, votre poids dans cinq années) ou d'expliquer avec une variable d'exposition d'intérêt (par exemple, est-ce que le traitement A est meilleur que le traitement B dans le contrôle de la douleur?) [2].

Les méthodes de sélection des variables automatiques

Les méthodes « datadriven » ne dépendent que des données et variables collectées. Elles ne reposent (en général) aucun *a priori*. Elles sont de différentes sortes. Les deux principaux sont les « test based » : l'apport de la variable dans le modèle est testé. Les « penalized » : les coefficients des variables sont réduits (ou pénalisés) [3]. Pour un même jeu de données, elles donnent des résultats différents. Par exemple, voici les variables sélectionnées selon quatre méthodes de sélection sur le jeu de données Prostate [4] du package R lasso2 (Tableau 1) :

Univarié	Backward	Elastic Net	Lasso
lcavol	lcavol	lcavol	lcavol
lweight	lweight	weight	lweight
-	age	age	-
-	lbph	lbph	lbph
svi	svi	svi	svi
lcp	-	lcp	-
gleason	-	gleason	-
pgg45	-	pgg45	pgg45

Tableau 1 : Variables sélectionnées selon la méthode

Les réflexions *a priori*

Inclure certaines variables peut complètement annuler voir inverser l'effet du traitement évalué. Pour bien choisir les variables qui vont corriger le biais sans annuler l'effet du traitement il faut passer par différentes étapes de construction du modèle. Il faut tenir compte de la chronologie, du sens des liens entre variables pour détecter les variables de médiation, confusion et collision. L'effet modérateur ou synergique entre deux variables est important à prendre en compte via les termes d'interaction. Par exemple, le fait de mettre de la crème solaire va modérer la dose d'UV reçue. Dans une étude observationnelle la plus part du temps les données ne sont pas recueillies à l'origine pour répondre à notre question de recherche précise (mais pour du soin courant ou la constitution d'un registre ou pour de la comptabilité). De ce fait certaines variables importantes sont non mesurées. Ces variables non disponibles pourront être substituées par des variables de substitution [5].

Les représentations a priori

Les graphes acycliques ou diagrammes causaux [6] (Figure 1) sont des représentations graphiques « qualitatives » qui aident à choisir les variables. Les modèles structuraux eux, ont pour vocation d'être inférentiels sur des relations complexes intriquées, directes et indirectes entre variables.

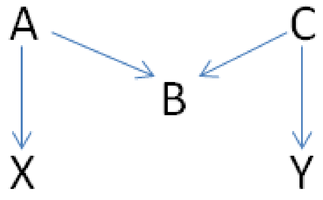

Figure 1: Diagramme de causalité (M-bias)

La reproductibilité de la recherche en médecine passe par la définition formelle des variables, et du savoir a priori. Retracer tous les liens pour chaque étude est parfois une perte de temps. La réutilisabilité de ces bases de données grandissantes fait que la définition des variables et les représentations pourraient être réutilisées. La taille de ces diagrammes causaux ou « brain maps » dépasse l'humain. L'utilisation de l'ontologie calquée sur ces représentations enrichies pourraient être une solution.

L'ontologie et la logique de description floue

De nouveaux outils tels que fuzzy PROTÉGÉ [7] permet de créer un modèle ontologique flou. Sa logique de description floue permet d'exprimer des faits avec d'autres descripteurs que « vrai » ou « faux. » Par exemple, la température corporelle (variable linguistique) a un référentiel (domaine de variation) et il existe un vocabulaire pour décrire celui-ci (par exemple : fièvre, hypothermie et valeurs « normales »). Les termes de ce vocabulaire est un sous ensemble flou. La fonction d'appartenance est moins stricte que la logique standard : on pourra dire que 39°C est de la fièvre mais qu'entre 37.8 et 38.5°C il y a un « flou ».

Objectifs

Le but est d'utiliser une ontologie à destination des biostatisticiens dans la recherche biomédicale pour la construction de modèles multivariés. Son opérationnalisation épaulera le biostatisticien en proposant des variables à inclure dans le modèle et d'autres à ne pas mettre en expliquant pourquoi.

Bibliographie

1. Moses LE. Measuring effects without randomized trials? Options, problems, challenges. Med Care. avr 1995;33(4 Suppl):AS8-14.

2. Shmueli G. To Explain or to Predict? Statist Sci. août 2010;25(3):289-310.

3. Desboulets L. A Review on Variable Selection in Regression Analysis. Econometrics. 23 nov 2018;6(4):45.

4. Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate : II. radical prostatectomy treated patients, Journal of Urology 141(5), 1076–1083.

5. Lubin JH, Hauptmann M, Blair A. Indirect adjustment of relative risks of an exposure with multiple categories for an unmeasured confounder. Annals of Epidemiology. nov 2018;28(11):801-7.

6. Pearl J. Causality: models, reasoning, and inference. 2. Cambridge: Cambridge University Press; 2009

7. Musen MA. The protégé project: a look back and a look forward. AI Matters. 16 juin 2015;1(4):4-12.

CHU ROUEN NORMANDIE

D2IM Département d'Informatique et d'Information Médicales

LIMICS