Fiche d'aide à la formalisation d'une question de modélisation

cf. lexique au verso pour la définition de chaque terme

Objectif principal de l'étude = question scientifique (si cela s'y prête sous forme de question **PICO** – Population / Intervention / Comparison / Outcome)

Quel est le cadre d'utilisation du modèle envisagé : **explicatif** ou **prédictif**, sur **quelle population** et quel **contexte cible** ?

Variable étudiée = **variable à expliquer =** ex. critère (outcome) primaire dans un essai clinique
c'est-à-dire **variable que l'on veut expliquer/prédire/modéliser** en fonction d'autres variables dites explicatives

Caractérisation statistique de cette variable : **qualitative** (si oui avec combien de modalités ? ATTENTION très compliquée si > 2), ou **quantitative continue** (a-t-on une idée de la distribution attendue ?), **quantitative censurée** (ex. temps de survie sachant que tous ne sont pas morts en fin d'étude), **discrète** (ex. score avec un grand nombre de modalités).

Variables explicatives =

variables principales dont on souhaite estimer l'effet sur la variable à expliquer (ex. le traitement dans un essai clinique) **+ variables concomitantes** que l'on doit prendre en compte car elles sont susceptibles d'avoir un effet sur la variable à expliquer, même si l'objectif principal n'est pas de modéliser leur effet (ex. la cage dans laquelle est hébergée l'animal, l'âge de l'animal, …)

Caractérisation statistique des variables explicatives. Une variable explicative qualitative est appelée un facteur et une variable explicative quantitative est appelée une covariable lorsqu'elle n'est pas une variable principale contrôlée.

Pour les facteurs, indication du nombre de modalités. Pour les variables explicatives quantitatives, indication de la relation attendue entre celle-ci et la variable à expliquer. Attend-on une relation monotone, sans effet seuil. Si non, envisager de la transformer en facteur en définissant des classes, car les modèles classiques supposent une relation linéaire entre chaque variable explicative quantitative et la variable à expliquer (hypothèse FORTE !).

Origine des données :

expérience envisagée, ou **méthode de recueil de données d'observation** (ex. dans une étude retrospective, méthode envisagée), en veillant à ce que les **critères d'inclusion** soient **cohérents** avec la **population** et le **contexte cible**. **Effectifs** attendus dans chaque groupe. Calcul de puissance possible ?

Facteur(s) aléatoire(s) à prendre en compte le cas échéant

ex. suivi de chaque animal au cours du temps -> facteur animal, animaux regroupés dans diverses cages -> facteur cage, données issus de plusieurs élevages -> facteur élevage

Type de modèle envisagé :

modèle linéaire, modèle linéaire mixte, modèle linéaire généralisé (le plus classique = régression logistique) modèle linéaire généralisé mixte, modèle de survie ?

Lexique

par ordre alphabétique

Covariable = variable explicative continue non contrôlée pendant la collecte de données (ex. l'âge des animaux si ce n'est pas la variable explicative principalement étudiée).

Facteur = variable explicative qualitative

Facteur aléatoire = facteur dont les valeurs dans les données peuvent être considérées comme un échantillon aléatoire à partir d'une plus grande population de valeurs. Dans un modèle mixte on prend en compte l'effet des facteurs aléatoires sur la variabilité (expliquée par chaque facteur) de la variable à expliquer (ex. la cage dans laquelle est hébergé l'animal, ou l'animal si plusieurs observations sont faites sur un même animal, …). Un facteur non aléatoire est appelé **facteur fixe**. Une variable explicative quantitative n'est jamais considérée comme aléatoire.

Modèle de régression logistique = modèle qui décrit une variable binaire (qualitative à deux modalités) en fonction de variables explicatives quantitatives et/ou qualitatives, en supposant un modèle d'erreur binomial et une relation linéaire entre le logit de la probabilité de l'événement modélisé ($logit(p) = ln\left(\frac{p}{1-p}\right)$, par ex. probabilité de guérir à l'issue du traitement, et les variables explicatives.

Modèle de survie = modèle qui décrit une variable de type « time-to-event », c'est-à-dire **temps d'atteinte d'un événement donné**, comme la mort par ex., en prenant en compte le fait que les **données** associées sont souvent **censurées**, du fait que tous les individus n'ont pas atteints cet événement en fin d'étude (-> information parfois partielle sur ce temps, par ex. > au temps passé par l'individu dans l'étude).

Modèle explicatif = modèle que l'on utilise pour mieux comprendre un phénomène biologique en caractérisant / quantifiant l'effet des variables explicatives sur la variable à expliquer.

Modèle linéaire = modèle linéaire gaussien = décrit une variable quantitative continue en fonction de variables explicatives quantitatives et/ou qualitatives, en supposant un modèle d'erreur additif gaussien et une relation linéaire entre la variable à expliquer et les variables explicatives.

Modèle linéaire mixte = modèle linéaire qui prend en compte l'effet d'un (ou de plusieurs) facteur(s) aléatoire(s).

Modèle prédictif = modèle que l'on souhaite utiliser pour prédire la variable étudiée en fonction des variables explicatives (par ex. pour établir un pronostic au moment du diagnostic d'une maladie).

Paramètre = un des coefficients d'un modèle, qui est estimé à partir des données (ex. la pente d'une droite de régression). Il convient donc d'éviter d'utiliser le terme paramètre pour désigner ce qu'on appelle en statistique une variable.

Variable à expliquer = variable dépendante = variable que l'on cherche à expliquer ou à prédire (celle que l'on représente sur l'axe des Y généralement).

Variable concomitante qui ne fait l'objet d'aucune manipulation de la part du chercheur, n'est généralement pas contrôlée, n'est pas une variable explicative principale, mais qui est susceptible de produire un effet de confusion d'effets si on ne la prend pas en compte. Si on la prend bien en compte dans le modèle, elle devient une variable explicative à part entière. Une variable concomitante peut être un facteur fixe (ex. le sexe de l'animal), un facteur aléatoire (ex. l'élevage dont il provient) ou une covariable (ex. l'âge de l'animal).

Variable(s) explicative(s) = variable(s) indépendante(s) = variables utilisées dans le but d'expliquer ou de prédire la variable dépendante (qu'on représente sur l'axe des X quand il n'y en a qu'une).