Projet de Biostatistiques

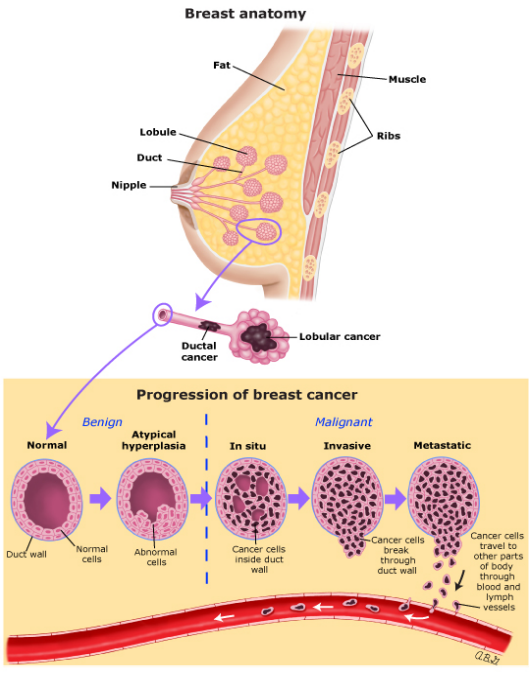
Le projet proposé porte sur l'analyse d'un jeu de données collectées aux États-Unis dans le cadre du programme SEER (Surveillance, Epidemiology and End Results) dirigé par l'Institut National du Cancer (NCI). Ce projet sera réalisé en groupe six groupes de quatre personnes et un groupe de trois. Dans une première partie, il est demandé de faire une analyse descriptive des données fournies. À partir de cette analyse, vous devrez définir une question d'intérêt sur cette problématique pouvant être résolue par une approche biostatistique et tenter d'y répondre. L'attendu est un compte-rendu écrit court (2 pages maximum) et une présentation orale de 8 minutes par groupe de travail, dans lesquels vous détaillerez la question d'intérêt que vous aurez choisie, votre démarche et vos résultats.

Contexte

Le cancer du sein

Le cancer est une maladie caractérisée par la prolifération incontrôlée de cellules, liée à un échappement aux mécanismes de régulation qui assure le développement harmonieux de notre organisme [1].En se multipliant de façon anarchique, les cellules cancéreuses donnent naissance à des **tumeurs** de plus en plus grosses qui se développent en envahissant puis détruisant les zones qui les entourent. La cellule cancéreuse qui se détache de la tumeur, peut migrer dans une autre partie du corps voisine voire un autre organe pour créer une nouvelle tumeur. Un des premiers signes de cette capacité invasive des cellules tumorales est leur présence dans les **ganglions lymphatiques** proches du lieu de la tumeur primaire.

Le cancer du sein est le cancer le plus fréquent chez la femme. Il représente plus du tiers de l'ensemble des nouveaux cas de cancer chez la femme en France. Seule une petite partie des cancers du sein, 5 à 10%, sont héréditaires, c'est-à-dire attribuable à une mutation génétique (qu'elle soit identifiée ou non).La recherche a permis d'identifier un certain nombre de **mutations génétiques** favorisant la survenue de cancers du sein chez les femmes. Le plus souvent, celles-ci portent sur des gènes appelés BRCA1 et BRCA2 (pour BReast Cancer 1/2 : gène 1/2 du cancer du sein). Il s'agit ici d'une predisposition genetique mais de multiples **facteurs de risques** favorisent l'apparition du cancer du sein [2].



Une particularité du cancer du sein est qu'il s'agit dans 60 à 70 % des cas d'un **cancer hormonodependant**. Les tumeurs hormonodépendantes se forment principalement dans des tissus dont le fonctionnement est normalement régulé par des hormones. Notamment chez les femmes atteintes d'un cancer du sein, les hormones sexuelles telles que **la progesterone et l'oestrogene** jouent un role dans le developpement et la progression de la tumeurs [3].

La prise en charge des patientes atteintes d'un cancer du sein se fait dans un premier temps par une etude diagnostiques visant à établir **l'étendue de la maladie, ou stade** (Grade ou Stage en anglais). Cette classification des patientes se fondent sur la taille de la tumeur, l'invasion des ganglions lymphatiques et la présence de nouvelles tumeurs dans d'autres organes (metastases). Une fois le stade determiné, des examens complementaires sont réalisés par exemple pour determiner si les cellules tumorales presentent à leur surface des recepteurs aux hormones rendant le cancer hormonodependant. Toutes ces informations rasssembleées permettent ensuite d'orienter le traitement du cancer vers les differentes solutions existantes [4].

Les données disponibles

Le jeu de données contient des informations de patientes atteintes de cancers du sein issues de la base de données du programme SEER de l'Institut national du cancer (NCI) aux Etats-Unis. Le NCI s'attache à fournir des statistiques sur le cancer afin de réduire son incidence et le fardeau qu'il represente pour la population et le systeme de santé.

Ce jeu de données repertorie particulierement des femmes atteintes d'un sous type de cancer du sein : des carcinomes lobulaires invasifs. Les données sont issues de la mise à jour du programme SEER de Novembre 2017 mais les patientes ont ete diagnostiquées entre 2006 et 2010.

Les variables rapportées dans le fichier (SEER_breast_cancer.csv) sont des informations sur la patiente (age, vivante ou decede) et des informations sur son cancer telles que la taille de la tumeurs, la positivité aux hormones et le grade du cancer en fonctions de plusieurs classifications. Vous trouverez la description de chaque variable dans le fichier SEER_breast_cancer_description.csv.

Étude des données

Lecture des fichiers

Le fichier SEER_breast_cancer.csv est au format CSV, comportant en première ligne le nom des colonnes. Le séparateur de colonne est la virgule. Les valeurs manquantes sont codées par une absence de caractères. Après lecture, vous devez vérifier que les données lues sont cohérentes avec celles présentes dans le fichier initial (nombre de lignes et colonnes lues, identité des valeurs dans les premières et dernières lignes, type ou classe des variables lues, ...).

Le fichier SEER_breast_cancer_description.csv, également au format CSV, comprend en première colonne le nom des variables (i.e. les colonnes du fichier SEER_breast_cancer.csv), en deuxième colonne leur description, et en troisième colonne des ressources pour plus de details.

Les questions suivantes ont pour objectif de vous guider dans l'analyse des données et de vous donner des exemples de questions à résoudre, mais l'évaluation portera sur une question en particulier que vous devrez définir et à laquelle vous devrez répondre à l'aide des données

Description des variables

1. De quel type sont les variables du jeu de données (qualitatives, quantitatives, discrètes, continues...) ?
2. Lesquelles contiennent des valeurs manquantes ? En quelles proportions ?
3. Calculez les informations statistiques (moyenne, médiane, écart-type) que l'on peut obtenir sur ces variables.
4. Représentez graphiquement la distribution de ces variables (e.g. histogrammes, boxplots)
5. Representez graphiquement la taille de la tumeur en fonction des autres variables telles le statut oestrogene, le N.Stage et l'age.
6. Certaines variables necessitent d'être recodées en catégorie. Proposer quelles variables et recoder les en categorie, puis representez les graphiquement en fonction d'autres variables.

Analyse

1. Selectionnez uniquement les patientes avec un T.Stage de 1 à 3. Evaluez le lien du T.Stage avec la taille de la tumeur. Que pouvez-vous dire sur la variable T.Stage ? Proposez des critères pour classer les patientes en T1,T2 ou T3.

2. Evaluez le lien entre la taille de la tumeur et le nombre de mois de survie. Vous pouvez faire de meme avec les variables que vous avez recodé précedemment.
3. Evaluez le lien entre l'age et le bilan hormonal.
4. Evaluez le lien entre la survie de la patiente et son bilan hormonal.
5. Calculez pour chaque patiente le taux de ganglions lymphatiques positifs.Evaluez le lien entre ce taux et l'age de la patiente ou avec la taille de la tumeur.
6. Plusieurs classifications des patientes sont disponibles dans ce jeu de données à cause d'une mise à jour du systeme. Representez la repartition des patientes dans ces differents systemes et en fonction de T et N. Que pouvez-vous en dire ? (*Vous pourrez utiliser la fonction facet_wrap pour faciliter les représentations*)
7. Quels facteurs de confusion peuvent influencer les liens étudiés ?

Evaluation

L'évaluation de votre travail se fera sur une seule analyse, vous devrez donc :

- choisir une question d'intérêt à laquelle il est possible de répondre via le jeu de données disponibles (vous pouvez vous inspirer de celles posées dans ce document, ou en proposer une nouvelle,
- poser les hypothèses,
- réaliser les tests statistiques nécessaires,
- produire les graphique pertinents illustant votre question d'intérêt,
- analyser les résultats et proposer une réponse,
- effectuer une synthèse.

Notes

1. Definition de la Ligue Contre le Cancer, https://www.ligue-cancer.net/article/26088-quest-ce-que-le-cancer
2. E-Cancer, Les principaux facteurs de risques, https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Facteurs-de-risque
3. Fondation ARC - Les cancers hormonodependants, https://www.fondation-arc.org/traitements-soins-cancer/hormonotherapie/quest-ce-quun-cancer-hormono-dependant
4. E-Cancer, Les traitements du cancer du sein, https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Traitements