BE Statistique & Santé

Cancer du sein : Facteurs pronostiques de son évolution à long terme

Ce projet est à effectuer par groupe de 2 étudiants.
Vous devrez me rendre, pour le vendredi 10 février 2017, un rapport (de 30 pages maximum, hors annexes). Vous pouvez effectuer ce projet avec les logiciels de votre choix (SAS, R, ...).

1 Problématique

Des chercheurs de l'institut Curie (Paris) ont mené une étude épidémiologique pour étudier le rôle pronostique des récepteurs hormonaux sur l'évolution de patientes qui ont eu un premier cancer du sein et sont en phase de surveillance post-thérapeutique.
Il y a deux types de récepteurs hormonaux présents en quantité plus ou moins grande dans la tumeur : les récepteurs oestrogéniques (R0) et les récepteurs progestéroniques (RP). On considère que les récepteurs sont présents dans la tumeur dès que la quantité mesurée dépasse le seuil de 10 fmol/mg.
L'évolution des patientes est mesurée par plusieurs variables qui constituent des données censurées. Quatre types d'événements peuvent apparaître suite à un premier épisode de cancer du sein :

- une récidive locale (c.a.d dans le même sein),
- un cancer controlatéral (c.a.d dans l'autre sein),
- des métastases,
- le décès (lié au cancer).

Pour résumer ces quatre événements, on peut étudier le disease-free interval correspondant au délai sans aucun évènement carcinologique (ni récidive locale, ni cancer controlatéral, ni métastase, ni décès lié au cancer). On le traduit en français par "délai de survie sans rechute".

2 Objectifs du projet

L'objectif général de ce projet est d'étudier les facteurs liés à la survenue d'un évènement suite à un premier épisode de cancer du sein.

Dans cette analyse, vous pourrez utiliser les méthodes d'analyse statistique que vous avez déjà étudiées en L3 et M1, et mettre en application de nouvelles méthodes que nous introduirons au cours du BE selon les besoins. Plusieurs analyses sont possibles. En effet, on dispose ici de données longitudinales et il existe de nombreux façons de les traiter et de les représenter.

Ce projet s'appuie sur des données réelles. La présence de valeurs manquantes est possible et vous devez les prendre en considération dans votre travail. De plus, vous pouvez (et devrez) également créer de nouvelles variables qui seraient appropriées pour votre analyse. Tout cela

devra être expliqué dans votre rapport.

Avant de commencer votre travail d'analyse statistique, réfléchissez à un plan précis de vos analyses et des méthodes à utiliser (selon la nature des variables). Ce plan et les méthodes d'analyse statistique utilisées devront être clairement décrits dans votre rapport. Ils doivent faire l'objet d'une partie "Matériel et Méthodes".
Dans cette partie, vous devrez également présenter les données et décrire les variables initiales ainsi que les nouvelles variables que vous avez dues créer pour votre analyse.

L'un des objectifs de ce projet est de montrer l'intérêt de prendre en compte l'aspect longitudinal des données dans les analyses statistiques. En effet, étudier la présence d'un évènement ou étudier la survenue d'un évènement au cours du temps ne s'analysent pas avec les mêmes méthodes, et peuvent donc donner des résultats différents et/ou complémentaires.

On vous propose un plan d'analyse statistique classique que vous pourrez ajuster et compléter selon vos idées :

1. **Description de la population étudiée**
   Décrire les caractéristiques des femmes et le type d'évènements survenus au cours du suivi.

2. **Traitement des données manquantes**
   Comme vous le verrez dans les premières analyses descriptives, vous serez confrontés ici à la présence (parfois importante) de données manquantes (ou éventuellement, aberrantes). Elles devront être prises en compte dans cette étude statistique (suppression des individus, nouvelle modalité, remplacement, ...). A la fin de cette étape, il est impératif que votre base de données soit figée et que toutes les analyses que vous menerez par la suite soient réalisées sur le même nombre d'observations.

3. **Relation entre les évènements et les caractéristiques des femmes**
   Des analyses bivariées et des modélisations pourront être mises en œuvre dans le but de mettre en évidence les caractéristiques liées à la survenue d'évènements.

4. **Modélisation longitudinale par des méthodes d'analyse de survie**
   Dans cette étude, on surveille la survenue d'un évènement carcinologique : s'il ne survient pas au cours du suivi, on sait que le délai de survenue de l'évènement est supérieur à la durée du suivi ; on dit que l'évènement est censuré. Des méthodes existent permettant d'analyser des données censurées en prenant en compte toute l'information disponible.
   La première difficulté de cette partie résidera dans la définition de la variable arrêt/censure et du délai correspondant.
   Dans un deuxième temps, vous pourrez étudier l'influence éventuelle des caractéristiques des patientes sur la survenue d'évènements carcinologiques au cours du temps.

3 Données

Les données que vous analyserez dans le cadre de ce projet concernent 2257 femmes suivies suite à un premier épisode de cancer du sein survenu entre 1974 et 1984. Les informations disponibles sur ces 2257 femmes portent sur leur situation au moment du premier cancer et sur l'évolution de leur état au cours du temps vis-à-vis de la maladie.

Le fichier de données `recepteurs` (format csv) contient les variables suivantes :

- `IDENT` : identifiant de la femme

Concernant les évènements survenus au cours du suivi :

- `E_DECES` : Présence (1) ou absence (0) du décès
- `E_META` : Présence (1) ou absence (0) de métastases
- `E_RECI` : Présence (1) ou absence (0) d'une récidive locale
- `E_CONT` : Présence (1) ou absence (0) d'un cancer controlatéral

Concernant les dates relevées au cours du suivi (au format jj/mm/aa) :

- `D_DN` : Date des dernières nouvelles
  Note : La date des dernières nouvelles est la date du dernier recueil d'informations. Elle a été fixée au 1er septembre 1993 pour les patientes encore vivantes et encore suivies. Pour les patientes décédées, il s'agit de la date de décès. Pour les patientes perdues de vue, il s'agit de la date à laquelle ont pu être relevées les dernières informations.
- `D_FIRST` : Date de survenue du premier cancer
- `D_DECES` : Date du décès s'il a eu lieu (valeur manquante notée NaN si le décès n'a pas eu lieu)
- `D_META` : Date de la survenue des métastases si elles sont apparues (valeur manquante notée NaN si cet évènement n'a pas eu lieu)
- `D_RECI` : Date de la survenue d'une récidive si elle a eu lieu (valeur manquante notée NaN si cet évènement n'a pas eu lieu)
- `D_CONT` : Date de la survenue d'un cancer controlatéral s'il a eu lieu (valeur manquante notée NaN si cet évènement n'a pas eu lieu)
  Note : il est possible que des évènements aient eu lieu mais qu'ils n'aient pas pu être datés.

Concernant la situation de la femme au moment du premier cancer :

- `AGE` : Age (en années)
- `MENOP` : Statut hormomal
  - 0 : pré-ménopause
  - 1 : post-ménopause

Concernant le type de cancer, la gravité et la prise en charge thérapeutique :

- `TNM` : Stade de gravité du cancer (classification proposée par l'Union Internationale Contre le Cancer, de 1 à 4)

- `TAILLE` : Taille de la tumeur primaire (en mm)
  Note : Dans les études, on utilise souvent des classes : moins de 20mm, entre 20 et 50mm, 50 mm ou plus.

- `ROQUANT` : Quantité de récepteurs œstrogéniques (RO) présents dans la tumeur (en fmol/mg)

- `RPQUANT` : Quantité de récepteurs progestéroniques (RP) présents dans la tumeur (en fmol/mg)
  Note : Comme dit précédemment, on considère que les RO et les RP sont présents si leur quantité est supérieure au seuil de positivité de 10 fmol/mg.

- `HISTO` : Type histologique de la tumeur
  - 1 : canalaire
  - 2 : lobulaire
  - 3 : autre

- `SBR` : Grade histologique de la tumeur de Scarff-Blomm et Richardson (de 1 à 3)

- `NBGANG` : Nombre de ganglions lymphatiques auxiliaires
  Note : on propose souvent dans les études de considérer des classes : 0, entre 1 et 3, 4 et plus.

- `CHIR` : Type de chirurgie effectuée sur le premier cancer
  - 0 : absence
  - 1 : tumorectomie
  - 2 : mastectomie

- `RAD` : Traitement par radiothérapie
  - 0 : absence
  - 1 : présence

- `CAUSEDC` : Cause du décès s'il a eu lieu
  - 1 : lié au cancer
  - 3 : non lié au cancer
  - valeur manquante notée NaN si le décès n'a pas eu lieu.