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

Max Halford - Master 2 SID

Mai 2017

Contents

1 *Introduction*

2 *Matériel et méthodes*

1. Statistiques récapitulatives des données mises à disposition
2. Préparation des données
   - Traitement des valeurs manquantes
   - Extraction de nouvelles variables
3. Analyses bivariées pour determiner les effets des caractéristiques sur les évènements
   - Test de Wilcoxon-Mann-Whitney pour les variables continues
   - Test du χ^2 suivi du calcul du V de Cramér pour les variables discrètes
4. Modélisation avec une régression logistique avec et sans sélection de variables
5. Modélisation avec une analyse de survie pour prendre en compte l'aspect longitudinal des données
6. Récapitulatif des deux modèles

3 *Analyse descriptive*

Les analyses suivantes se font en ignorant les valeurs manquantes qui seront traitées par la suite.

On a à disposition un jeu de données qui concerne 2257 femmes ayant eu un premier épisode de cancer du sein entre 1974 et 1984. Après le premier épisode, chaque femme a été suivi et on dispose d'un suivi individuel qui peut aller jusqu'au 1er septembre 1993 (dans le cas où la patiente est encore vivante et suivie). Lors du suivi, 4 types d'évènements différents ont été enregistrés:

- *Décés*: la patiente est morte, que ce soit à cause du cancer ou pas.
- *Métastase*: un cancer du sein est dit métastatique lorsque des cellules cancéreuses issues de la tumeur initiale se sont installées dans un autre organe du corps comme par exemple au niveau des os, des poumons ou du foie.
- *Récidive*: un nouvel épisode cancéreux a eu lieu dans le même sein que lors de l'épisode initial.
- *Cancer controlatéral*: le cancer s'est propagé à l'autre sein.

682 des 2257 (30%) patientes sont décédées au cours de leur suivi; il se peut aussi que certaines des patientes perdues de vue soit décédées sans qu'on ne le sache.
Il va de soit que ces évènements ne sont pas indépendants, d'ailleurs en regardant le tableau suivant on s'apercoit que les évènements de décès et de métastases sont liés.

Table 1: Co-occurences des évènements (effectifs)

	E_DECES	E_META	E_RECI	E_CONT
E_DECES	682	428	135	32
E_META	428	589	151	33
E_RECI	135	151	307	18
E_CONT	32	33	18	105

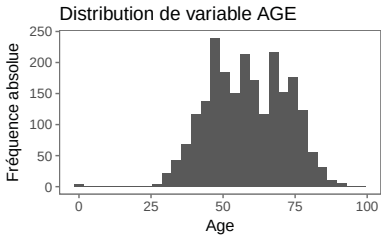
Table 2: Co-occurences des évènements (fréquences, les lignes somment à 1)

	E_DECES	E_META	E_RECI	E_CONT
E_DECES	0.53	0.34	0.11	0.03
E_META	0.36	0.49	0.13	0.03
E_RECI	0.22	0.25	0.50	0.03
E_CONT	0.17	0.18	0.10	0.56

Ici on a transposé la matrice des évènements et on l'a multiplié par elle-même ($X^t X$) pour obtenir les co-occurences d'évènements deux à deux.

On constate que 428 des 589 (73%) patientes faisant signe d'une métastate sont décédées; de plus 428 des 682 (63%) patientes décédées étaient atteintes d'une métastase. En d'autres termes le risque de décès pour les patientes atteintes d'une métastate est 2.5 fois plus élevé. Les évènements n'apparaissent pas à la même fréquence (les éléments diagonaux représentent le nombre d'occurences de chaque évènement); cela peut être du au fait qu'un évènement en enclenche un autre mais seulement dans un sens.
On dispose aussi de la date d'occurence des évènements, on pourra donc par exemple étudier l'ordre d'apparition des évènements au cours du temps ou bien la prévalence du cancer au cours du temps. Lors de la préparation des données il faudra accorder du temps à la manipulation de ces dates, notamment en les convertissant dans un format analysable.

Le risque relatif (RR) de décès en cas de métastate est calculé de la façon suivante: $\frac{P_M(D)}{P_{\bar{M}}(D)} = \frac{0.726}{0.290} = 2.5$



Le reste des variables a été mesuré lors de l'épisode initial du cancer. Lors de cette épisode, les patientes ont en moyennes 58 ans et 1466 (65%) d'entre elles sont ménopausées.
La majeure partie (92%) des cancers des patientes ont été initiallement classifiés comme étant au stade 1 ou 2 (respectivement 34% et 58%) selon la classification de l'UICC [1].

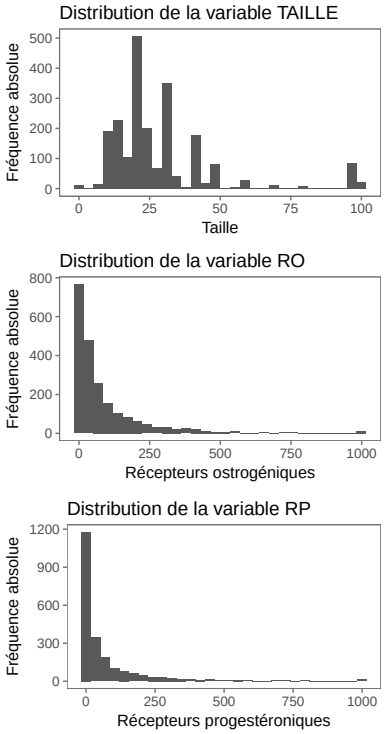
[1] UICC: Union Internationale Contre le Cancer.

Table 3: Stade de gravité du cancer

	1	2	3	4
Effectif	753	1296	131	56

	1	2	3	4
Proportion	0.34	0.58	0.06	0.03

Quand à la taille de la tumeur primaire, sa distribution fait signe d'une disymétrie à gauche avec une moyenne de 28.9 millimètres. Cette disymétrie vient tout simplement du fait que la taille minimale est de 0.



Les mesures de récepteurs œstrogéniques (RO) et de récepteurs progestéroniques (RP) présents dans la tumeur initiale semblents suivre des distributions exponentielles décroissantes. La mesure RO a une valeur médiane de 37 et une moyenne de 92.65 alors que la mesure RP a une valeur médiane de 14 pour une moyenne de 73.22. Si on opère une binarisation de ces quantités de récepteurs en considérant dans les deux cas un seuil de positivité strict de 10 fmol/mg, on obtient les fréquences suivantes. On peut dire que le RO est jugé présent dans 1556 (72%) des cas tandis que le RP l'est dans 1194 (54%) des cas.

Table 4: Présence du récepteur RO

	<= 10	> 10
Effectif	611	1556
Proportion	0.28	0.72

Table 5: Présence du récepteur RP

	<= 10	> 10
Effectif	1016	1194
Proportion	0.46	0.54

Pour ce qui est du stade histologique de la tumeur de Scarff-Blomm-Richardson (noté SBR), il atteint modalités 3 dont les fréquences sont données dans le tableau suivant. Il y'a 3 stades possibles, le premier étant le meilleur pour la patiente.

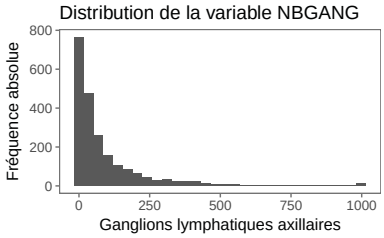
Table 6: Mesure SBR

	1	2	3
Effectif	466	1144	402
Proportion	0.23	0.57	0.20

De même que pour les quantités de RO et de RP, le nombre de ganglions lymphatiques axillaires (et non pas auxilaires) semble aussi être distribué de façon exponentielle. Ces ganglions peuvent augmenter de volume chez la femme en cas de cancer du sein. Si on effectue un découpage pour donner plus de sens "humain" à cette quantité, on obtient les fréquences suivantes. On constate que le fait d'avoir plus de 3 ganglions est rare puisque cela représente seulement 5% des patientes.

Table 7: Nombre de ganglions lymphatiques axillaires

	Aucun	Entre 1 et 3	Plus de 3
Effectif	1517	443	101
Proportion	0.74	0.21	0.05



La fréquence de chaque type de chirurgie effectuée sur le premier épisode (abscence, tumorectomie ou mastectomie) est reprèsentè dans le tableau suivant. On constate que 1824 (81%) des patientes ont reçu une chirurgie.

Table 8: Type de chirurgie au premier cancer

	Abscence	Tumorectomie	Mastectomie
Effectif	429	727	1097
Proportion	0.19	0.32	0.49

La variable RAD indique si la patiente a poursuivi un traitement par radiothérapie. Il y'en quasiment autant qui en ont suivi qu'il y'en a qui n'en ont pas suivi.

Table 9: Traitement par radiothérapie

	Non	Oui
Effectif	1146	1104
Proportion	0.51	0.49

Enfin, 537 des 637 (79.7%) des patientes décédées le sont à cause du cancer; 413 de ces 537 (76.9%) patientes faisaient signe d'une métastate.

Table 10: Raison du décès et lien avec la métastate

	Décès cancer	Décès autre
Métastate	124	128
Pas de métastate	413	8

4 *Préparation des données*

Comme il se doit, le jeu de données à disposition étant un jeu de données réel, il contient des valeurs manquantes et abhérrantes.

4.1 Gestion des données manquantes

Tous les évenements n'ont bien heureusement pas de valeurs manquantes. Pour ce qui est des variables explicatives, toutes en ont hélas au moins quelques unes comme on peut le voir sur le graphique suivant. Les lignes sont des observations et les parties blanches correspondent à des valeurs manquantes.

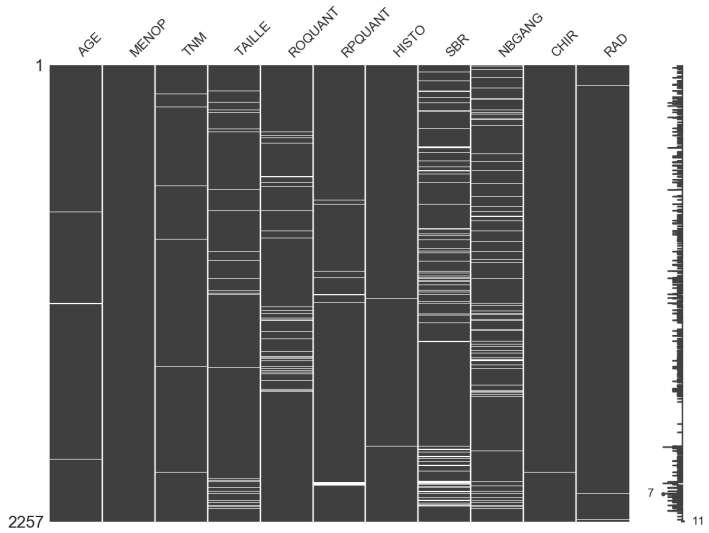
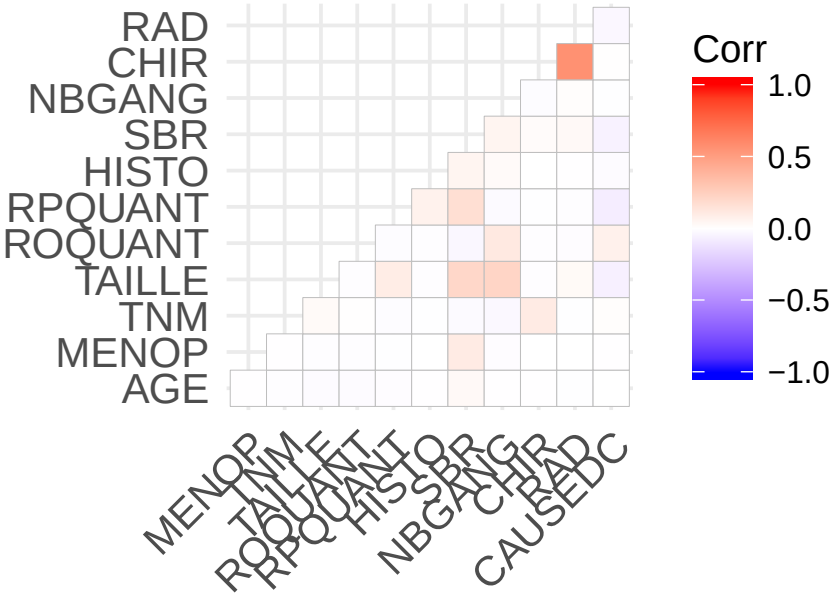


Figure 1: Matrice des valeurs manquantes

Il ne paraît pas y avoir de corrélations entre la présence de valeurs manquantes des différentes variables. On peut confirmer cela avec une carte de chaleur qui montre la corrélation des dites valeurs manquantes. Il semble cependant y avoir un lien pour les variables RAD et CHIR, toutefois c'est probablement seulement dû au hasard puisque il n'y a que 7 valeurs nulles pour la variable RAD et 4 pour la variable CHIR.



Pour ce qui est du nombre de valeurs manquantes par ligne, en voici ci-dessous un tableau récapitulatif.

Table 11: Distribution du nombre de valeurs manquantes par observation

	0	1	2	3	4
Effectif	1696	430	104	24	3
Proportion	0.75	0.19	0.05	0.01	0.00

On constate qu'il y'a seulement 27 (1%) d'observations qui ont plus de deux valeurs manquantes et 131 (6%) qui ont plus d'une. Une majorité de 1696 (75%) observations n'ont pas de valeurs manquantes et 430 (19%). En retirant les observations qui ont plus d'une valeur manquante on garderait 94% des données; de plus en faisant cela les variables MENOP et CHIR n'ont plus de valeurs manquantes. Il y'a aussi 4 patientes dont l'âge indiqué est de 0, on les retire du jeu de données, de cette façon il n'y plus de valeurs manquantes pour la variable AGE. Enfin il y'a une observation qui n'a pas de date de dernières nouvelles, on l'enlève.
Nous avons conservé 2111 (93.6%) des données. On peut maintenant à remplacer les valeurs manquantes. La variable TNM qui représente le stade de gravité du cancer n'a que 6 (~0%) valeurs manquantes et c'est une variable discrète, on peut donc tout simplement remplacer les valeurs manquantes par le mode de la distribution qui est 2. On peut appliquer ce même processus pour les variables HISTO, SBR et RAD qui sont aussi des variables discrètes et ont respectivement 2, 2

et 3 valeurs manquantes.
Le reste des variables avec des valeurs manquantes sont toutes des variables continues avec la caractéristique qu'elles font toutes signe d'une disymétrie. On va donc remplacer les valeurs manquantes de chaque par sa médiane respective pour éviter de trop prendre en compter les valeurs extrêmes. Ceci concerne 4 variables qui sont TAILLE, ROQUANT, RPQUANT et NBGANG et qui ont respectivement 29 (1.3%), 59 (2.7%), 22 (1%), 124 (5.8%) valeurs manquantes.

4.2 Extraction de nouvelles variables

Avant de développer des modèles on va extraire de nouvelles variables. On va aussi remanier le jeu de données pour le faciliter les modélisations qui vont suivre.
Tout d'abord on définit l'évènement "disease-free" comme étant le cas où aucun des évènements indiqués n'a lieu. Il y'a 1161 (55%) des observations où c'est le cas.
En prochaine étape on "applatit" (en anglais *melt*) le jeu de données pour avoir une vision plus "chronologique" des données. Cela est plus simple à comprendre en comparant les deux tableaux suivants.

Table 12: Partie du jeu de donneés initial

IDENT	E_DECES	E_META	D_DECES	D_META	ROQUANT
1	1	1	07/01/90	18/01/90	22
2	0	0	NULL	NULL	9
3	1	0	18/10/83	NULL	0

Table 13: Version "applatie" du tableau précédent

IDENT	EVENT	DATE	ROQUANT
1	E_DECES	07/01/90	22
1	E_META	18/01/90	22
3	E_DECES	18/10/83	0

En applatissant le jeu de données on se rend qu'il y'a d'autres données manquantes. En effets certains évènements n'ont pas de date associée (4), on retire ces évènements du jeu de données. De plus quelques évènements de décès n'ont pas de cause associée (8), on remplace les valeurs manquantes par la modalité la plus fréquente qui est 1 (indiquant que le décès est liė au cancer).

Maintenant que le jeu données possède une colonne DATE, il est trivial de calculer la différence en jours entre la date du premier évènement cancéreux et la date d'occurrence des divers évènements. On peut aussi extraire l'année de chaque date, ceci sera possiblement informatif puisque on peut supposer que la qualité des traitements reçus s'améliore avec le temps. Enfin, même si on a applatit le jeu de données on peut converser les occurrences d'évènements en indiquant si oui (1) ou non (0) un évènement antécédant a eu lieu. Concrètement on aura trois variables booléennes indiquant si oui ou un certain type d'évènement a eu lieu avant l'évènement observé; ces trois colonnes sont E_META, E_RECI et E_CONT. Il n'y a pas besoin de garder E_DECES puisque étant le dernier évènement qui a lieu il ne peut pas aider à expliquer l'occurrence d'autres évènements, au contraire du reste des évènements.
Au final on a 18 variables explicatives qui sont en ordre alphabétique:

- AGE
- CAUSEDC (qui ne peut que être utilisé dans le cas de l'évènement de décès)
- CHIR
- D_FIRST_YEAR
- DATE_YEAR
- DIFF_JOURS
- E_CONT
- E_META
- E_RECI
- HISTO
- MENOP
- NBGANG
- RAD
- ROQUANT
- RPQUANT
- SBR
- TAILLE
- TNM

A côté de ça nous avons 3 colonnes qui serviront à contruire, possiblement, des modèles séparées. Celles-ci sont:

- IDENT
- DATE
- EVENT