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

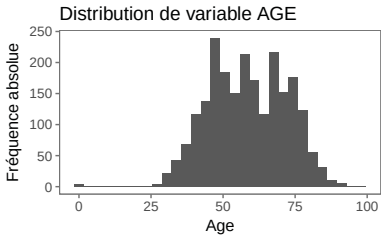
Table 2: Co-occurences des évènements (fréquences, les lignes somment à 1)

	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].

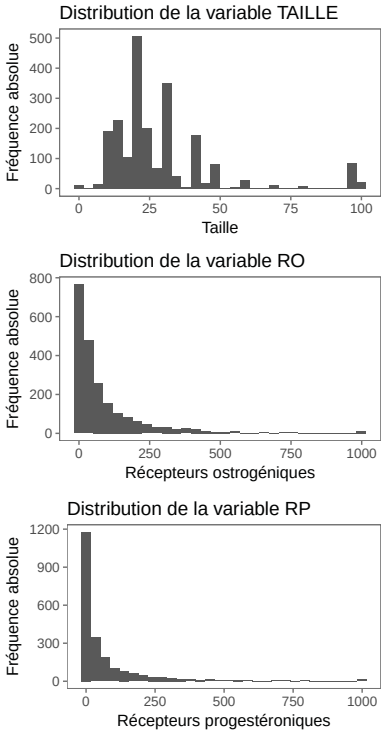
[1] UICC: Union Internationale Contre le Cancer.

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.

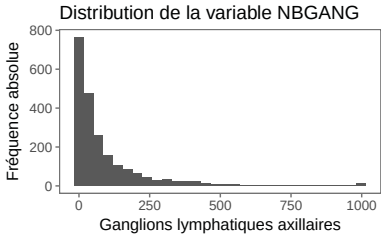
Table 6: Mesure SBR

	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.

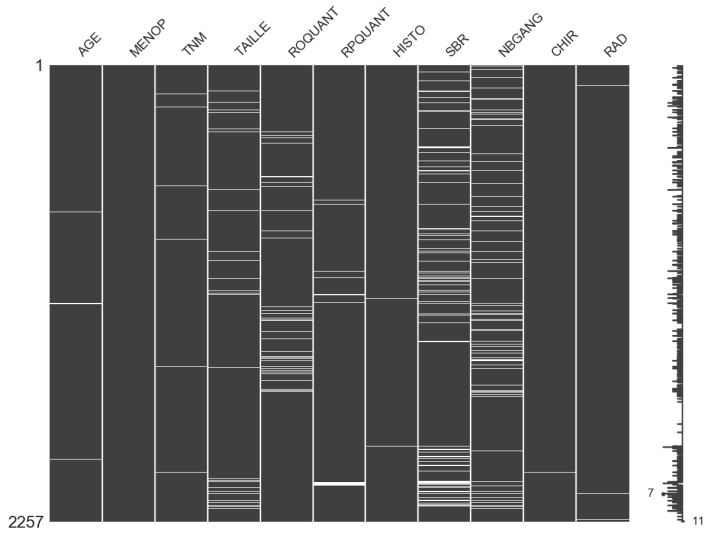
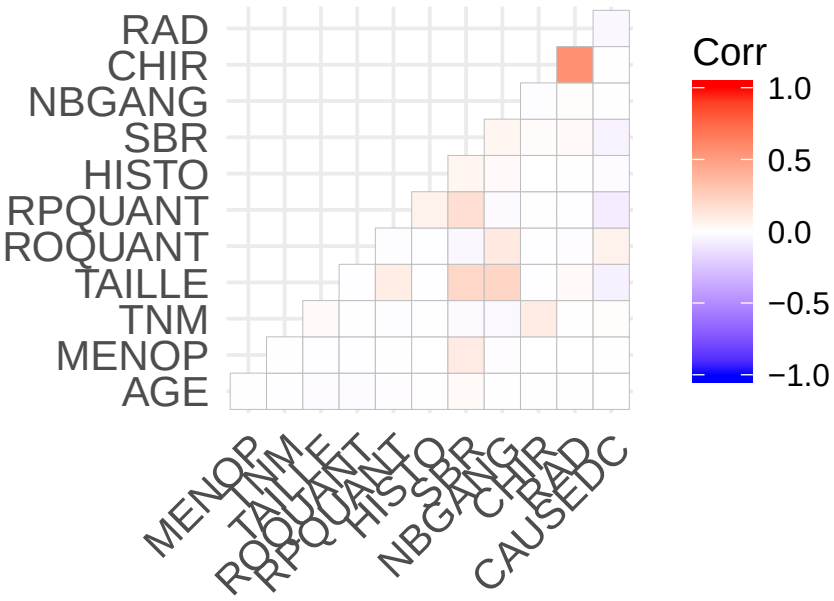


Figure 1: Matrice des valeurs manquantes

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%) observations. 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	D_DN	ROQUANT
1	1	1	07/01/90	18/01/90	18/01/90	22
2	0	0	NULL	NULL	16/01/90	19
3	1	0	18/10/83	NULL	18/10/83	0

Table 13: Version "applatie" du tableau précédent

IDENT	EVENEMENT	DATE	OCCURENCE	ROQUANT
1	E_DECES	07/01/90	1	22
1	E_META	18/01/90	1	22
1	E_DECES	16/01/90	0	19
1	E_META	16/01/90	0	19
3	E_DECES	18/10/83	1	0
3	E_META	18/10/83	0	0

Pour ce qui est des évènements on en distingue 7:

- E_DECES_CANCER: la patiente est décédée à cause du cancer
- E_DECES_AUTRE: la patiente est décédée pour une raison autre

que le cancer
- E_META: apparition de métastates
- E_RECI: récidive locale
- E_CONT: cancer controlatéral
- E_DF: la patiente a fait d'aucun évènement listé précedemment
- E_SURVIE: la patiente a survécu à la data d'observation

En applatissant le jeu de données on se rend que 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 comporte 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'èvènement 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 17 variables explicatives:

- AGE
- 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
- EVENEMENT

Enfin la variable réponse OCCURRENCE permet de savoir si ou non un évènement a eu lieu.

5 *Statistiques bivariées*

On peut d'abord commencer à résumer l'influence de chaque variable sur l'occurrence d'un évènement, et ceci pour chaque évènement. Pour les 10 variables continues on peut calculer la p-valeur donné par le test de Wilcoxon-Mann-Whitney; l'avantage d'utiliser ce test est qu'il est non-paramétrique et donc qu'on a pas à, par exemple, supposer que nos données proviennent d'une distribution normale. En ce qui concerne les 7 variables catégoriques on peut effectuer un simple test du chi^2 pour déterminer la présence de "lien"; on peut ensuite calculer le V de Cramér pour déterminer l'intensité de ce lien. Evidemment le V de Cramér a seulement du sens si le test du χ^2 est significatif. Le V de Cramér varie de 0 à 1 et exprime une certaine corrélation entre deux variable catégoriques.

5.1 *Variables continues*

Table 14: p-valeurs des tests de Wilcoxon-Mann-Whitney

	Décès cancer	Décès autre	Méta	Reci	Cont	Disease-free	Survie
AGE	0.000	0.000	0.002	0.000	0.002	0.000	0.000
TNM	0.000	0.005	0.000	0.932	0.001	0.000	0.000
TAILLE	0.000	0.542	0.000	0.003	0.667	0.000	0.000
ROQUANT	0.012	0.729	0.254	0.100	0.676	0.872	0.026
RPQUANT	0.000	0.147	0.001	0.970	0.058	0.020	0.000
SBR	0.000	0.943	0.000	0.002	0.277	0.000	0.000
NBGANG	0.000	0.534	0.000	0.041	0.871	0.000	0.000
DATE_YEAR	0.000	0.000	0.000	0.000	0.000	0.000	0.000
D_FIRST_YEAR	0.000	0.000	0.000	0.942	0.943	0.000	0.000
DIFF_JOURS	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Il semble que toutes les variables continues aient un effect significatif sur le décès lié au cancer. Le tableau suivant montre les moyennes de chaque variable selon que le décès lié au cancer ait eu lieu (1) ou pas

(0). Les p-valeurs disponibles ci-dessus indiquent si ces moyennes sont significativement différentes ou pas.

Table 15: Moyennes des variables continues pour les décès liés au cancer

	0	1
AGE	58.18	60.73
TNM	1.72	1.92
TAILLE	27.77	32.83
ROQUANT	91.83	95.56
RPQUANT	78.30	58.27
SBR	1.91	2.19
NBGANG	0.47	1.46
DATE_YEAR	1989.85	1984.85
D_FIRST_YEAR	1980.99	1979.49
DIFF_JOURS	3264.75	1955.03

Les patientes décèdées sont en moyenne 30 mois plus âgées lors de l'apparition du cancer. Comme on peut s'y attendre, e stade de leur cancer est plus élevé, la taille tumeur primaire est plus élevée d'un demi-centimètre. Quand au récepteurs, la quantité RO est plus élevée alors que la quantité RP est plus basse. Le grade histologique de la tumeur est plus élevé et le nombre de ganglions est trois plus élevé. Enfin les patientes sont mortes en moyenne 5 années après la survenue du premier épisode.
Pour ce qui est des décès qui ne sont pas liés au cancer, les variables relatives au premier cancer ne sont pas significatives (c'est à dire les variables TAILLE, ROQUANT, RPQUANT, SBR et NBGANG). On voit que les variables liés aux années sont encore très significatives, ceci est du au fait que lorsqu'un un patient meurt un évènement est enregistré. La variable DATE_YEAR et DIFF_JOURS est très explicative puisque toutes les patientes ont commencé à être suivies la première année, cependant comme cette variable n'est pas disponible lors de l'arrivée du premier cancer on ne peut pas s'en servir pour faire de la prédiction. La variable DIFF_JOURS est tout simplement la différence entre DATE_YEAR et D_FIRST_YEAR, puisque D_FIRST_YEAR reste stable DIFF_JOURS aura donc la même puissance explicative que DATE_YEAR. En apprentissage machine on dit que les variables DATE_YEAR et DIFF_JOURS "leak" (en français "fuient") de l'information. Tout dépend de comment est posée la question à laquelle on veut répondre; si la date de l'évènement alors on peut utiliser cette information puisque les évènements de décès arrivent plus vite que la moyenne, si dans

l'autre cas on nous demande la probabilité de décès lors de la survenue du premier cancer on ne peut pas utiliser ces variables.

Table 16: Moyennes des variables continues pour les décès non liés au cancer

	0	1
AGE	57.88	73.28
TNM	1.76	1.94
TAILLE	28.98	28.25
ROQUANT	91.69	108.99
RPQUANT	73.95	68.84
SBR	1.97	1.98
NBGANG	0.69	0.83
DATE_YEAR	1988.95	1984.55
D_FIRST_YEAR	1980.72	1979.30
DIFF_JOURS	3025.83	1910.31

En ce qui concerne l'apparition de métastates, mis à part la variables ROQUANT, toutes les autres semblent être liés. Il se peut que ceci soit du au fait que la métastate appairaisse en même temps que le cancer et que par enchaînement toutes les variables soient liées aux deux types d'évènements. D'autant plus que sur le tableau suivant on constate que les variations de moyennes observées pour les cancers sont les mêmes que pour la métastase. Seulement la variable AGE a une moyenne plus faible.

Table 17: Moyennes des variables continues pour les métastates

	0	1
AGE	59.37	57.07
TNM	1.73	1.86
TAILLE	27.43	33.25
ROQUANT	90.05	100.21
RPQUANT	77.26	63.41
SBR	1.91	2.16
NBGANG	0.48	1.32
DATE_YEAR	1989.70	1983.67
D_FIRST_YEAR	1980.94	1979.80
DIFF_JOURS	3225.50	1411.38

Pour ce qui est de la récidive, seulement les variables AGE, TAILLE, SBR, DATE_YEAR et DIFF_JOURS semblent être explicatives. Les

patientes atteintes d'un premier cancer à un âge ont plus de chance d'avoir une récidive, de même que les patientes avec une large première tumeur.

Table 18: Moyennes des variables continues pour les récidives

	0	1
AGE	59.50	53.95
TNM	1.77	1.76
TAILLE	28.26	33.51
ROQUANT	94.29	82.05
RPQUANT	73.33	75.82
SBR	1.96	2.08
NBGANG	0.65	1.04
DATE_YEAR	1988.76	1984.41
D_FIRST_YEAR	1980.64	1980.63
DIFF_JOURS	2983.70	1365.13

Les cancers controlatéraux, quand à eux, semblent survenir dès que le stade du cancer est plus faible, que l'âge est bas et la quantité de récepteurs progestéroniques est élevée.

Table 19: Moyennes des variables continues pour les cancers controlatéraux

	0	1
AGE	58.96	54.76
TNM	1.78	1.56
TAILLE	28.82	31.62
ROQUANT	94.10	62.12
RPQUANT	73.39	79.52
SBR	1.97	2.04
NBGANG	0.69	0.80
DATE_YEAR	1988.67	1985.56
D_FIRST_YEAR	1980.64	1980.63
DIFF_JOURS	2953.65	1783.67

Toutes les variables sauf ROQUANT semblents affecter l'évènement "disease-free".

Table 20: Moyennes des variables continues pour l'évènement disease-free

	0	1
AGE	59.84	57.90
TNM	1.85	1.70
TAILLE	31.77	26.63
ROQUANT	98.76	87.73
RPQUANT	69.55	77.02
SBR	2.08	1.88
NBGANG	1.07	0.39
DATE_YEAR	1986.54	1990.45
D_FIRST_YEAR	1979.99	1981.17
DIFF_JOURS	2401.29	3419.65

De même que pour les évènements de décès liés au cancer, toutes les variables sont liées à la survie des patientes.

Table 21: Moyennes des variables continues pour la survie

	0	1
AGE	63.04	57.00
TNM	1.93	1.70
TAILLE	31.76	27.77
ROQUANT	97.87	90.54
RPQUANT	60.62	79.07
SBR	2.14	1.90
NBGANG	1.32	0.44
DATE_YEAR	1984.79	1990.31
D_FIRST_YEAR	1979.45	1981.14
DIFF_JOURS	1949.44	3381.19

5.2 *Variables catégoriques*

Table 22: p-valeurs des tests du χ^2

	Décès cancer	Décès autre	Méta	Reci	Cont	Disease-free	Survie
MENOP	0.000	0.000	1.000	0.000	0.002	0.006	0.000
HISTO	0.295	0.259	0.040	0.004	0.856	0.002	0.074
CHIR	0.000	0.000	0.001	0.245	0.052	0.000	0.000
RAD	0.179	0.000	0.177	0.717	0.011	0.166	0.000
E_META	0.000	0.000	NaN	0.423	0.921	0.000	0.000

	Décès cancer	Décès autre	Méta	Reci	Cont	Disease-free	Survie
E_RECI	0.000	0.424	0.000	NaN	0.009	0.000	0.000
E_CONT	0.000	0.821	0.010	1.000	NaN	0.000	0.003

Table 23: Vs de Cramér

	Décès cancer	Décès autre	Méta	Reci	Cont	Disease-free	Survie
MENOP	0.078	0.141	0.000	0.105	0.068	0.060	0.141
HISTO	0.034	0.036	0.055	0.073	0.012	0.078	0.050
CHIR	0.126	0.148	0.079	0.036	0.053	0.146	0.196
RAD	0.029	0.099	0.029	0.008	0.055	0.030	0.078
E_META	0.572	0.085	NaN	0.017	0.002	0.465	0.497
E_RECI	0.185	0.017	0.174	NaN	0.057	0.258	0.163
E_CONT	0.078	0.005	0.056	0.000	NaN	0.131	0.064

La ménopause semble être significativement lié au décès dus au cancer, en effet des 489 patientes mortes à causes du cancer, 352 (71.98%) ont atteint la ménopause alors que 1022 des 1622 (63%) patientes qui ne sont pas mortes à cause du cancer l'ont atteinte. En lieu de faire le détail ici, on pourra résumer l'importance des variables avec les odds ratios obtenus avec la régression logistique dans la partie suivante.

6 Régression logistique

On peut procéder à une régression logistique pour estimer la puissance prédictive de nos variables. De plus, on peut faire cela pour chacun des types d'évènements dont on dispose. Pour obtenir une estimation qui soit un tant soi peu proche du monde réel, on peut effectuer une validation croisée avec 10 plis. Pour estimer la performance des différentes logistiques mises en place on peut utiliser l'aire sous la courbe ROC (qu'on dénote ROC AUC) puisque celle-ci mesure la capacité à ordonner les observations selon la probabilité d'obtenir un 1, ceci reflète un cas réel qui serait de prioriser les soins patientes selon leur probabilité de mourir (en termes hospitaliers faire du *triage*).
Si on veut prédire le fait de décéder à cause du cancer, on obtient une ROC AUC de moyenne 0.852 avec un écart-type insignifiant. En calculant l'exponentielle de chaque coefficient dans la régression logistique on obtient les odds ratios.

Table 24: Odds ratios pour prédire la survie des patientes

	Odds ratio
(Intercept)	0.0000000
AGE	0.9505916
MENOP1	1.5821518
TNM	0.8497294
TAILLE	1.0075117
ROQUANT	1.0003390
RPQUANT	1.0005605
HISTO2	1.8397533
HISTO3	1.6251635
SBR	0.6776103
NBGANG	0.9403224
CHIR1	0.4392441
CHIR2	0.1616225
RAD1	0.8201197
E_META1	0.0794649
E_RECI1	0.6371726
E_CONT1	0.7972943
DATE_YEAR	0.6917248
D_FIRST_YEAR	1.5479573
DIFF_JOURS	1.0019481

On constate que le fait d'avoir eu une métastate permet beaucoup de prédire si une patiente va survivre ou pas. En effet, 1438 des 1492 (96%) patientes qui ont survécu n'ont pas eu de métastate. En effet il faut faire attention en interprétant le tableau précédent: on prédit la probabilité de survivre et donc avoir un odds ratio inférieur à 1 veut dire que plus variable est élevée, plus la chance de survie augmente. Par exemple pour l'âge de la patiente, chaque année en plus réduit les chances de survie d'environ 5%.

7 Analyse de survie

7.1 Objectif

On veut maintenant étudier l'arrivée d'évènements, notamment la mort des patientes, au cours du temps. Le format du jeu de données actuel ne nécessite pas de changements puisqu'il indique, pour chaque observation, le type d'évènement, l'occurrence ou non et la date d'observation.

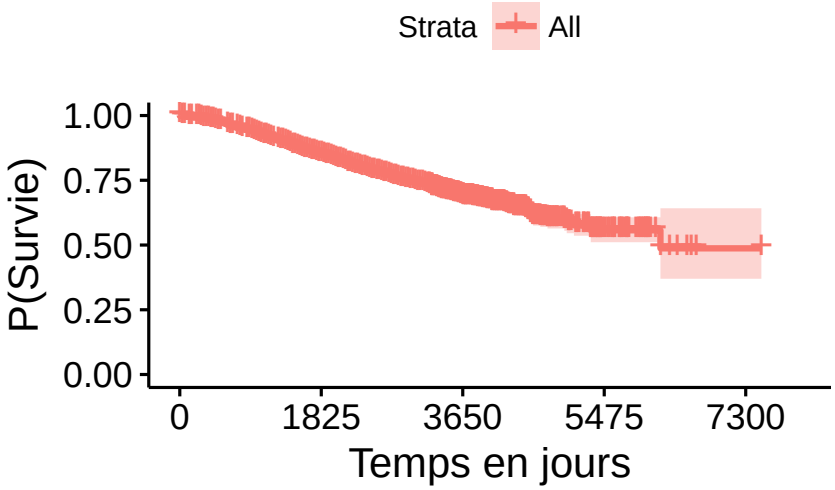
7.2 *Estimation de la survie globale*

Puisqu'on nos données sont censurées (à droite), on peut estimer la probabilité de mourir au cours du temps avec l'estimateur de Kaplan-Meier. Dans le cas où on connaitrait la date de décès de chaque patiente on aurait simplement à calculer une fonction de survie de façon déterministe. On peut afficher la proportion de personnes en vie au cours dans un tableau.

Table 25: Evolution du taux de survie au cours du temps

Jours	Taux de survie	Erreur standard	IC 95% inf	IC 95% sup
0	0.9990526	0.0006702	0.9977410	1.0000000
775	0.9492781	0.0050569	0.9399160	0.9587335
1302	0.8996260	0.0073402	0.8867761	0.9126620
1798	0.8505990	0.0092529	0.8353122	0.8661656
2325	0.7998299	0.0111495	0.7825412	0.8175006
2914	0.7499005	0.0131220	0.7308600	0.7694372
3472	0.7016026	0.0154706	0.6806480	0.7232023
4247	0.6481811	0.0204047	0.6227702	0.6746288
4743	0.5979600	0.0300175	0.5637950	0.6341953
5301	0.5568150	0.0440828	0.5107255	0.6070638
6200	0.4872132	0.1407140	0.3697796	0.6419409

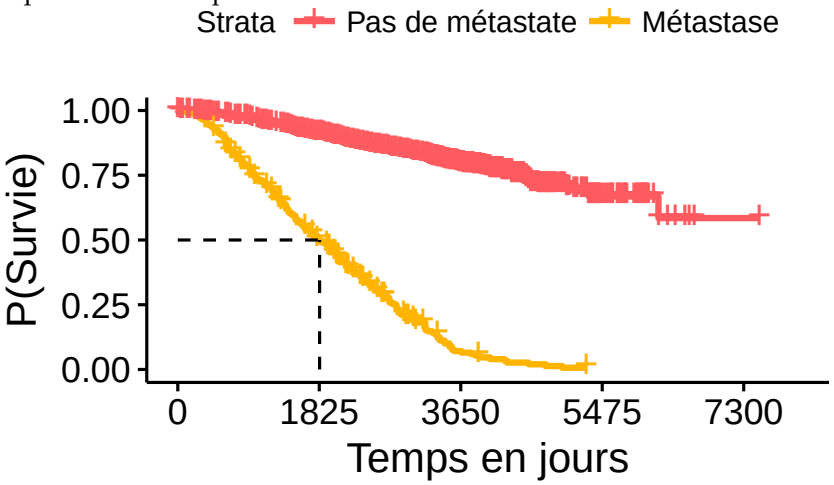
On constate que cela un moins de 10 ans pour que 30% des patientes soient mortes. On peut aussi représenter cette évolution à l'aide d'une courbe de survie.



On constate bien que les patientes ne meurent pas brutalement. Plus on avance dans le temps et moins l'estimation du taux de survie est fiable, ceci étant du à la réduction du nombre d'observations.

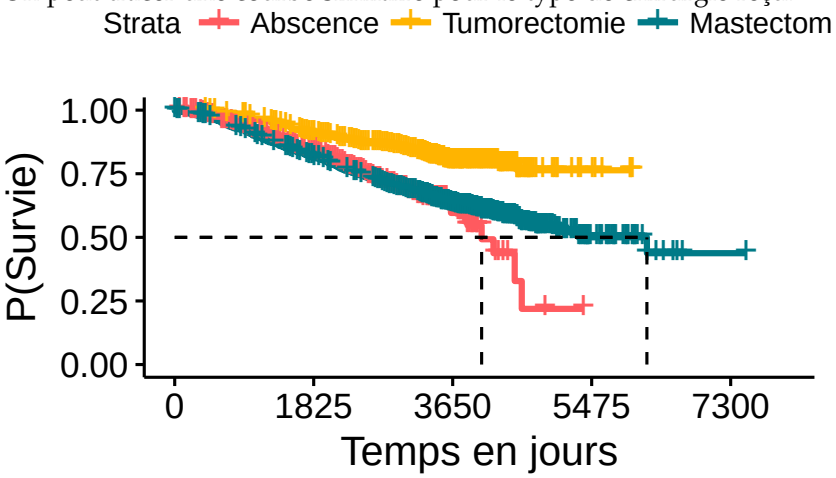
7.3 Estimation de la survie suivant un facteur

On peut estimer la probabilité de survie des patientes au cours du temps dans plusieurs groupes disjoints. On peut définir ces groupes à partir d'une ou plusieurs variables discrètes.



On voit bien que la probabilité de survie décroît beaucoup plus au cours du temps si une patiente a eu une métastate. En effet, seulement 50% des patientes qui ont une métastate survivent au bout de 5 années, alors que 90% des patientes qui ne sont pas atteintes d'une métastate survivent.

On peut tracer une courbe similaire pour le type de chirurgie reçu.



Ici on voit que les patientes qui se sont vu retirer leur tumeur via une tumorectomie ont une plus grande chance de survivre au cours. La mastectomie ne semble pas avoir d'effet sur la probabilité de survie avant 10 années, cependant après les patientes ayant eu une mastectomie survivent mieux.

7.4 Tests log-rank

A vu d'oeil on peut plus ou moins deviner si une variable a un effet sur le taux de survie. Une démarche plus rigoureuse consiste à faire recours à test statistique pour conclure. Le test log-rank permet justement de comparer des fonctions de survie de façon non-paramétrique. Intuitivement l'idée est de comparer le nombre d'évènements attendus à chaque pas de temps t entre les différents groupes formés à partir d'une variable. L'hypothèse nulle est que la probabilité de survie dans chaque groupe est la même à chaque pas de temps t.
Si on prend en compte la présence antérieure d'une métastase, on rejette l'hypothèse nulle et on conclut ce qu'on a vu graphiquement. On peut effectuer ce test pour chaque variable catégorique et garder celles pour lesquelles on rejette l'hypothèse nulle avec $\alpha = 0.05$. Il se trouve que le test est seulement rejeté pour la variable HISTO où la p-valeur est de 0.09.

8 Conclusion

Durant la phase de pré-traitement nous avons géré les valeurs manquantes et nous avons créé de nouvelles variables, notamment temporelles. De plus, nous avons normalisé les données pour pouvoir étudier l'effet des variables sur différentes variables de façon aisée. Ce travail préliminaire a permis de gagner beaucoup de temps par la suite.
En premier lieu nous avons examiner la distribution des variable de façon univariée, de plus nous avons pris le temps de mesurer la co-occurrence des différents types d'évènements. Ensuite nous avons effectué des analyses bivariées à travers des tests statistiques pour pouvoir mesurer l'impact des variables sur les différents types d'évènements. Nous avons aussi effectué une régression logistique pour obtenir les odds-ratios associés à chaque variable.
Enfin, l'analyse de survie a confirmé les observations faites grâce aux tests précédents. L'avantage certain de cette approche est de pouvoir chiffrer le taux de survie au fur et à mesure du temps. Cependant, il est moins pratique d'inclure plusieurs variables comme on l'aurait fait dans une régression sur le nombre de jours jusqu'à la mort. De plus, la prise en compte de variable continues nécessite de discrétiser ces dites variables et donc de perdre un peu d'information.