Projet MTS

Arno Barrabès

Projet MTS

Contexte

Les maladies transmises sexuellement (MTS) sont des maladies sournoises autant par les préjugés qui gravitent autour que par les mécanismes de transmission. Le dépistage des MTS n'est pas une question d'arsenal médical mais de stratégie.
Ainsi la difficulté ne réside pas dans les méthodes de diagnostique mais dans les moyens d'atteindre certains groupes cibles. Les grands complexes hospitaliers se sont résignés à jouer un rôle de second plan au profit des petites cliniques de quartier. Malgré tout, certaines personnes semblent rétissantes à se diriger vers les cliniques spécialisées. Plusieurs préfèrent le contact plus personnel avec un médecin de famille. Dans le but d'atteindre les groupes à haut risque, on a créé un programme de formation à l'intention des médecins de pratique privée. Ce programme permet aussi à ces médecins d'utiliser sans frais les laboratoires d'analyse des grands centres hospitaliers.

Contexte

Le médecin peut donc diagnostiquer directement une MTS sans devoir référer le patient à un centre spécialisé. Pour atteindre un groupe cible, il suffit de former des médecins opérant dans le milieu que fréquente ce groupe. Dans le but de mieux définir les groupes cibles, on doit réussir à pointer les facteurs de risque. On a, pour quelques médecins intégrés au programme, recensé tous les patients examinés dans le cadre du programme de dépistage. L'étude se limite essentiellement à la gonorrhée. Ce choix est imposé par une raison d'ordre pratique : la gonorrhée est la seule MTS commune qui est dépisté efficacement par un simple test de laboratoire sur une culture. Le recensement tient compte des informations sur le sexe, l'âge, l'orientation sexuelle, le nombre de MTS antérieures, la raison de la visite, le diagnostic ainsi que sur le nombre de partenaires différents dans le mois qui a précédé la visite.

Contexte

La littérature existante sur le sujet suggère la formation de quelques catégories :

- l'âge partage les gens en deux groupes : moins de 30 ans et 30 ans et plus.
- le nombre de MTS antérieures est une variable qui devrait être de style dichotomique (déjà ou pas encore).
- le nombre de partenaires peut, lui aussi, être considéré comme une variable dichotomique séparant les peu actifs des très actifs.

Contenu du rapport

L'objectif du rapport est donc de déterminer les groupes cibles qui présentent un grand risque de contracter des MTS. Nous disposons des données concernant 3144 sujets et 12 variables. Pour ce faire, nous avons effectué l'analyse statistique1 que nous vous présentons dans ce rapport comme suit :

1. Analyse descriptive
   - Description des variables
   - Analyse des fréquences et statistiques pertinentes
   - Tableaux croisés et liens entre les variables explicatives et la variable diagnostique
2. Ajustement par un modèle de régression logistique
   - Détermination du meilleur sous-ensemble de variables explicatives
   - L'ajustement
   - Caractérisation des groupes à risque et probabilités respectives
3. Conclusion
4. Annexes 1, 2, 3, 4
5. Références

Description des variables

Nous possédons les données recueillies sur 3144 sujets et regroupées dans les variables suivantes :

1. Sexe

1	:	homme
2	:	femme

2. État Civil

1	:	célibataire
2	:	marié
3	:	séparé / divorcé
4	:	veuf
5	:	pas de réponse

Description des variables

3. Age

1	:	Moins de 30 ans
2	:	30 ans et plus

4. Orientation sexuelle

1	:	homosexuel(le)
2	:	hétérosexuel(le)

5. MTS antérieures

1	:	non
2	:	oui

6. Nombre de MTS antérieures

1	:	aucun
2	:	au moins un

Description des variables

7. Raison de la visite

1	:	symptôme
2	:	contact
3	:	dépistage
4	:	contrôle
5	:	autre

8. Nombre de partenaires

1	:	0 ou 1 partenaire
2	:	2 ou plus

9. Histoire de contact

1	:	relation avec partenaire(s) contaminé(s) par MTS
0	:	pas de relation

Description des variables

10. Culture

0	:	négatif
1	:	gorge (positif)
2	:	col ou urètre (positif)
3	:	anus (positif)
4	:	gorge et urètre ou col (positif)
5	:	gorge et anus (positif)
6	:	urètre ou col et anus (positif)
7	:	les 3 sites (positif)

11. Diagnostique

0	:	pas de gonorrhée
1	:	gonorrhée

Analyse des fréquences et statistiques pertinentes

Dans cette section, nous nous référons aux tableaux et résultats de l'Annexe 1 : Fréquences etc . . .(pages 7 à 11).

Remarque

Note : les pourcentages donnés plus bas sont calculés après avoir enlevé les valeurs manquantes et sont arrondis.

- Tout d'abord le nombre total de sujets est de 3144.
- Pour la variable sexe, nous obtenons 77% d'hommes contre 23% de femmes. Il n'y a aucune valeur manquante.
- Pour la variable état civil, il y a 87% de célibataires. Il y a 4 valeurs manquantes.
- L'âge des sujets varie entre 14 et 78 ans, la moyenne est de 28 ans, l'écart type est de 8 ans, avec 62% des sujets âgés de moins de 30 ans. Il y a 6 valeurs manquantes.
- Pour l'orientation sexuelle, on retrouve 50% d'homosexuels contre 42% d'hétérosexuels. Il y a 257 valeurs manquantes.

Analyse des fréquences et statistiques pertinentes

- La variable MTS antérieures indique que 52% des sujets ont eu au moins une fois une MTS. Il n'y a aucune valeur manquante.
- Le nombre de MTS antérieures varie entre 0 et 65, la moyenne est de 1.2 avec un écart type de 2.56. On note que 13.5% des sujets ont eu 3 MTS ou plus dans le passé. Il n'y a aucune valeur manquante.
- La variable raison de la visite indique notamment que 47% des sujets éprouvaient des symptômes et 32% venaient pour un test de dépistage. Il y a 2 valeurs manquantes.
- Le nombre de partenaires pendant le dernier mois varie entre 0 et 98, la moyenne est de 3.2 et l'écart type est de 6.4. Dans l'échantillon, 35% ont 0 ou 1 partenaire tandis que 63% en ont eu plus de 1. Il y a 65 valeurs manquantes.

Analyse des fréquences et statistiques pertinentes

- La variable histoire de contact nous indique que 79% des sujets déclarent ne pas avoir eu de relation avec un partenaire contaminé, contre 19% qui eux déclarent avoir eu des relations à risque. Il y a 106 valeurs manquantes.
- La variable culture nous indique 73% des sujets ne sont pas atteints de gonorrhé, 24% le sont et 22.9% des sujets atteints ont été diagnostiqué ainsi par un prélèvement de la gorge. Il y a 72 valeurs manquantes.
- La variable diagnostique nous indique que 73% des sujets ne sont pas atteints de gonorrhé contre 24% qui le sont. Il y a 72 valeurs manquantes.

Sélection des variables

- Tel que mentionné par le client, la variable culture ne sera pas d'intérêt ici puisque nous nous intéressons au diagnostique, peu importe la méthode du prélèvement.
- De plus, puisque les variables MTS antérieures et nombre de MTS antérieures nous procurent essentiellement la même information, nous allons uniquement considérer la variable MTS antérieures pour la suite de l'analyse.

Remarque

Nous allons voir d'autres méthodes pour la sélection des variables.

Codage des variables

- En ce qui a trait aux variables âge et nombre de partenaires, de par la littérature existante sur le sujet, nous avons formé les catégories suivantes : « moins de 30 ans » et « 30 ans et plus » pour la variable âge. Nous ferons dorénavant référence à cette nouvelle variable sous le nom d'âge : variable catégorielle.
- La littérature conseille également de considérer la variable nombre de partenaires sous forme dichotomique en séparant les « peu actifs » des « très actifs ». Nous avons ainsi défini les sujets ayant eu 0 ou 1 partenaire le dernier mois comme étant peu actifs. Nous ferons dorénavant référence à cette nouvelle variable sous le nom de nombre de partenaires : variable catégorielle.

Sélection des variables

Les 9 variables qui feront parties de l'analyse sont donc :

1. Sexe
2. État Civil
3. Age : variable catégorielle moins de 30 ans / 30 ans et plus
4. Orientation sexuelle
5. MTS antérieures
6. Raison de la visite
7. Nombre de partenaires : variable catégorielle 0 ou 1 / 2 ou plus
8. Histoire de contact
9. Diagnostique

Tableaux croisés

Dans cette section, nous nous référons aux tableaux et résultats de l'annexe 2 : tableaux croisés (pages 12 à 16).

- La variable que l'on tente d'expliquer dans cette étude est le diagnostique et ce à l'aide des 8 autres variables explicatives.
- Il est donc d'intérêt de jeter un premier coup d'œil sur l'influence des variables explicatives sur le diagnostique.

Nous obtenons alors, pour chacune des variables explicatives, les tableaux croisés et les tests du khi-deux. Les résultats des tests du khi-deux présentés en annexe sont résumés dans le tableau suivant :

Tableaux croisés

Les résultats des tests du khi-deux présentés en annexe sont résumés dans le tableau suivant :

Variables explicatives	**Valeurs-p**
Sexe	0
État civil	0
Age : variable catégorielle	0.0025
Orientation sexuelle	0
MTS antérieures	0.067
Raison de la visite	0.26
Nombre de partenaires : variable catégorielle	0
Histoire de contact	0.624

Ainsi, on voit que les variables qui semblent le plus influencer le diagnostique sont : ***Sexe***, ***État civil***, ***Age : variable catégorielle***, ***Orientation sexuelle***, et ***Nombre de partenaires : variable catégorielle***. Elles ont toutes des valeurs-p inférieures à 0.05.

Ajustement par un modèle de régression logistique

À présent, nous désirons construire un modèle nous permettant de cibler les groupes d'individus présentant ***les plus hauts risques*** d'être atteints de la gonorrhée. Nous désirons donc connaître la ***probabilité*** qu'un individu de caractéristiques X possède la gonorrhée. Pour ce faire, nous allons prédire la variable diagnostique, à l'aide d'un ajustement par un modèle de régression logistique. Si l'on note par p : la probabilité d'un diagnostique positif pour un individu possédant les caractéristiques $X_1, X_2, \ldots, X_p$, alors un ajustement par la régression logistique donne :

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \quad [1]$$

1. Hosmer, D.W. & Lemeshov, S. (2000), Applied logistic regression, second edition Wiley

Détermination du meilleur sous-ensemble de variables explicatives

Dans cette section, nous nous référons au tableaux et résultats de l'annexe 3 : Sélection des variables explicatives (page 17) : l'analyse s'est faite sur 2664 sujets (voir le Tableau 0).
Afin de déterminer, parmi les 8 variables explicatives à notre disposition, le meilleur sous-ensemble de ces variables qui devrait être utilisé pour décrire le diagnostique dans un contexte de régression logistique :
Nous avons utilisé l'approche « forward-backward stepwise ». Le critère d'entrée des variables est fixé à 5% et celui de la sortie à 10%. À l'aide de cette méthode, les variables retenues sont les suivantes : ***sexe***, ***âge : variable catégorielle***, ***orientation sexuelle***, et ***nombre de partenaires : variable catégorielle***.

Remarque

On remarque que ce sont les mêmes variables, à l'exception de l'état civil, qui furent significatives pour les tests du khi-deux de la page 18.

L'ajustement

Remarque

*Dans cette section, nous nous référons au tableaux et résultats de l'annexe 4 : L'ajustement (page 18) : l'analyse s'est faite sur 2664 sujets : (voir tableau 0)

Le modèle ajusté est le suivant :

$$\begin{aligned} \log \frac{p(x)}{1-p(x)} &= -1.774 + 0.277 \times \text{SEXE} - 0.417 \times \text{AGE} \\ &+ 0.749 \times \text{ORIENT SEX} + 0.513 \times \text{NB PART} \end{aligned}$$

La présence de chacune des variables est significative (voir tableau 1 de l'annexe 4). Bien que la constante ne soit pas choisie dans les modèles de selection, nous avons jugé préférable de la conserver dans le modèle afin d'éviter d'obtenir un modèle trop simpliste.

L'ajustement

On trouve les ratios de cotes estimées « odds-ratio » (voir tableau 1 de l'annexe 4 et voir [2] pour plus d'explications). Le tableau ci-dessous les résume :

Variables	**odds-ratio estimés**
Sexe	1.32
Age : variable catégorielle	0.66
Orientation sexuelle	2.11
Nombre de partenaires : variable catégorielle	1.67

De par le dernier tableau, nous constatons que :

- Les hommes sont près d'une fois et demi plus à risque que les femmes.
- Les moins de trente ans sont une fois et demi plus à risque que les trente ans et plus.
- Les homosexuels sont plus de deux fois plus à risque que les hétérosexuels.
- Les personnes très actives (2 ou plus) sont près d'une fois et demi plus à risque que les personnes moins actives.

2. Neter, J., Kutner, M., Nachtsheim, C., Wasserman, W., Applied Linear Statistical Models, McGraw Hill, 1996.

Caractérisation des groupes à risque et probabilités respectives

Nous présentons le tableau résumant les caractéristiques de chacun des groupes d'individus concerné par notre modèle et leur probabilité de contraction respective (obtenues à partir de l'ajustement du modèle sur 2758 sujets) :

Caractéristiques	**Probabilités**
Homme, moins de trente ans, homosexuel, 2 partenaires ou plus	0.442
Femme, trente ans ou plus, hétérosexuelle, 0 ou 1 partenaire	0.101

On observe que d'après notre modèle utilisé, le groupe le plus à risque est caractérisé par les gens de sexe masculin, de moins de trente ans, homosexuel, ayant eu 2 partenaires ou plus le dernier mois. À l'opposé se trouve les gens de sexe féminin de trente ans ou plus, hétérosexuelle, ayant eu 0 ou 1 partenaire le dernier mois.

Conclusion

- Ainsi, nous avons fourni une description des données afin de cerner l'échantillon : fréquences, moyennes, valeurs manquantes, etc.
- Nous étions alors en mesure d'identifier clairement les variables d'intérêt pour l'étude.
- Ensuite, pour nous donner une idée des variables explicatives influençant le diagnostique, nous avons obtenu à l'aide des tableaux croisés et du test du khi-deux : Sexe, État civil, Age : variable catégorielle, Orientation sexuelle, et Nombre de partenaires : variable catégorielle.
- L'étape suivante fut de déterminer le meilleur sous ensemble des variables explicatives qui devrait être utilisé pour décrire le diagnostique dans un contexte de régression logistique.

Conclusion

- Les variables retenues furent : Sexe, Age : variable catégorielle, Orientation sexuelle, et Nombre de partenaires : variable catégorielle, soit les mêmes variables obtenues par les tests du khi-deux à l'exception de la variable état civil.
- Nous avons donc ajusté le modèle aux données, puis à l'aide de ce dernier, nous avons obtenu les ratios de cotes estimées.
- Ceci nous a permis de comparer les risques associés aux différents niveaux de chacune des variables du modèle.
- Enfin, toujours à l'aide du modèle, nous avons présenté les probabilités de contraction de la gonorrhée associées à chacun des groupes d'individus concerné par notre modèle.

Fréquences, statistiques, etc

Variable ***SEXE*** (1-homme, 2-femme) :

	counts	percentage
1	2429	77.25827
2	715	22.74173

Variable ***ETAT_C*** (1-célibataire, 2-marié, 3-séparé ou divorcé, 4-veuf) :

	counts	percentage
1.0	2742	87.324841
2.0	224	7.133758
3.0	167	5.318471
4.0	7	0.222930

Variable ***AGE*** :

count	3138.000000
mean	28.448693
std	7.828457
min	14.000000
25%	23.000000
50%	27.000000
75%	32.000000
max	78.000000

Variable ***AGE*** (0-moins de 30 ans, 1-plus de 30 ans) :

	counts	percentage
0.0	1953	62.237094
1.0	1185	37.762906

Tableaux croisés

DIAGN*SEXE (1-homme, 0-femme) :

DIAGN	0.0	1.0	All
SEXE			
0	533	90	623
1	1443	598	2041
All	1976	688	2664

Test de khi-deux entre les deux variables :

```
chi_val, p_val
(54.19741354942164, 1.8132536563726 45e-13)
```

DIAGN*RAISON (1-symptôme, 2-contact, 3-dépistage) :

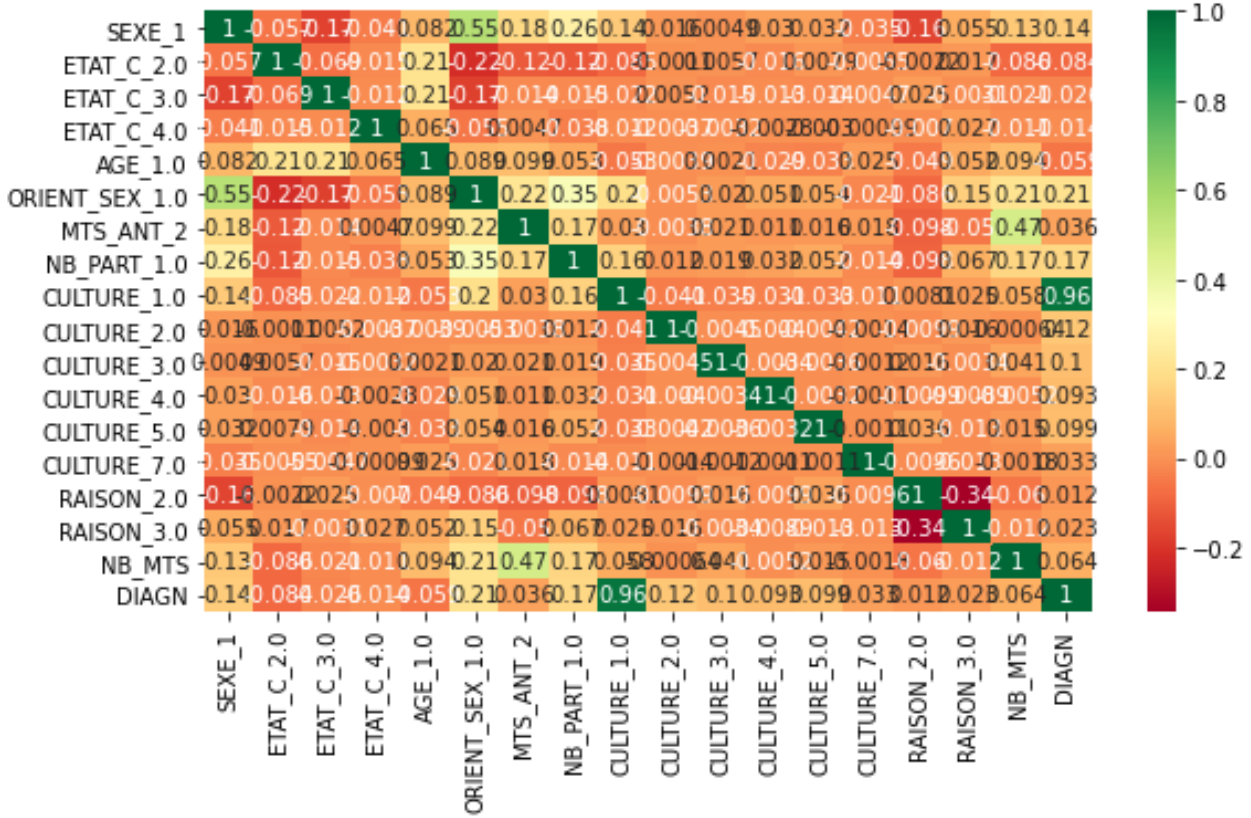
DIAGN	0.0	1.0	All
RAISON			
1.0	961	310	1271
2.0	384	141	525
3.0	631	237	868
All	1976	688	2664

Test de khi-deux entre les deux variables :

```
chi_val, p_val
(2.6491588694229513, 0.26591477009737297)
```

Sélection des variables explicatives

On peut afficher la matrice de corrélation, mais pas très clair ce qu'on observe.



Sélection des variables explicatives

On peut calculer le facteur d'inflation de variance (VIF) qui mesure la gravité de la multicolinéarité dans l'analyse de régression

	variables	VIF
0	SEXE_1	4.427740
1	ETAT_C_2.0	1.234078
2	ETAT_C_3.0	1.172617
3	ETAT_C_4.0	1.016435
4	AGE_1.0	1.888588
5	ORIENT_SEX_1.0	3.786955
6	MTS_ANT_2	2.587232
7	NB_PART_1.0	1.821140
8	CULTURE_1.0	inf
9	CULTURE_2.0	inf
10	CULTURE_3.0	inf
11	CULTURE_4.0	inf
12	CULTURE_5.0	inf
13	CULTURE_7.0	inf
14	RAISON_2.0	1.211036
15	RAISON_3.0	1.535208
16	NB_MTS	1.640084
17	DIAGN	inf

Du tableau ci-haut, on peut déduire qu'il existe la multicolinéarité (VIF>5). Par exemple, on peut enlever la variable ***CULTURE***.

L'ajustement

Le modèle de regression logistique :

Dep. Variable:	DIAGN	No. Observations:	2664
Model:	Logit	Df Residuals:	2659
Method:	MLE	Df Model:	4
Date:	Thu, 14 Oct 2021	Pseudo R-squ.:	0.05500
Time:	14:09:06	Log-Likelihood:	-1438.0
converged:	True	LL-Null:	-1521.7
Covariance Type:	nonrobust	LLR p-value:	3.779e-35

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7744	0.118	-15.033	0.000	-2.006	-1.543
SEXE_1	0.2777	0.150	1.857	0.063	-0.015	0.571
AGE_1.0	-0.4171	0.096	-4.326	0.000	-0.606	-0.228
ORIENT_SEX_1.0	0.7498	0.118	6.376	0.000	0.519	0.980
NB_PART_1.0	0.5136	0.097	5.274	0.000	0.323	0.704