Cancer Cervical risk analysis

Author : Andréa Boniffay

1. Introduction :

Les statistiques peuvent être des outils intéressants dans le domaine des sciences médicales pour aider à la compréhension et à la prise de décision. La régression logistique est l'une des méthodes statistiques les plus utilisées (1). Cette méthode fait partie de la famille des modèles statistiques appelés *Generalised Linear Models* (GLM). La régression logistique se distingue par le type de variable dépendante qu'elle traite, à savoir des résultats qualitatifs (2).

Le cancer du col de l'utérus est l'un des types de cancer les plus courants chez les femmes, en particulier dans les pays en développement. En 2020, le GLOBOCAN estimait la prévalence à 600 000 cas dans le monde et à plus de 300 000 décès (3). Le dépistage précoce et la compréhension des facteurs de risque sont essentiels pour améliorer les résultats en matière de santé publique.

La base de données utilisée dans ce rapport, fournie par Fernandes et al. (2017) (4), contient des informations sur divers facteurs de risque pour le cancer du col de l'utérus, notamment des antécédents médicaux, des comportements de santé et des variables cliniques. Ce projet traite de l'exploration des facteurs de risque du cancer cervical par l'utilisation des statistiques descriptives et de la *régression logistique*.

2. Materiels et Methodes

2.1. Description des données :

L'ensemble des données comprend des informations démographiques et antécédents médicaux de 858 patientes. Le dataset couvre un échantillon de patientes ayant fréquenté le service de gynécologie de l'hôpital universitaire de Caracas, au Venezuela de 2012 à 2013 (5). L'âge des patientes s'étend de 13 à 84 ans (moyenne = 27 ans). Toutes les patientes sont sexuellement actives et 98 % d'entre elles ont été enceintes au moins une fois. Les patientes sont issues de la classe socio-économique la plus basse et appartiennent donc à la population ayant le risque le plus élevé (6).

La base de données contient 36 variables, dont 24 sont des variables qualitatives binaires et 12 sont des variables quantitatives. Les variables incluent : âge (*années*), nombre de partenaires sexuels , âge lors du premier rapport sexuel (*années*), nombre de grossesses, consommation de tabac (*booléen*), années de consommation de tabac (*années*), paquets de cigarettes consommés par an, utilisation de contraceptifs hormonaux (*booléen*), durée d'utilisation des contraceptifs hormonaux (*années*), utilisation de dispositifs intra-utérins (*booléen*), durée d'utilisation des dispositifs intra-utérins (*années*), maladies sexuellement transmissibles (*booléen*), nombre de MST contractées, antécédents de condylomes (*booléen*), condylomes cervicaux (*booléen*), condylomes vaginaux (*booléen*), condylomes vulvo-périnéaux (*booléen*), syphilis (*booléen*), maladie inflammatoire pelvienne (*booléen*), herpès génital (*booléen*), molluscum contagiosum (*booléen*), SIDA (*booléen*), VIH (*booléen*), hépatite B (*booléen*), infection par le virus du papillome humain (*booléen*), nombre de diagnostics de MST, temps écoulé depuis le premier diagnostic, temps écoulé depuis le dernier diagnostic, diagnostic de cancer (*booléen*), diagnostic de néoplasie intraépithéliale cervicale (*booléen*), diagnostic de HPV (*booléen*), autres diagnostics (*booléen*), test de Hinselmann (*booléen*),test de Schiller (*booléen*), cytologie (*booléen*), biopsie (*booléen*).

Aucune variable n'est vide. Plusieurs patientes de la cohorte ont exprimé le souhait de ne pas répondre à certaines questions pour des raisons personnelles (7). 3 622 observations sur un total de 30 888 (11,7 %) sont manquantes. Parmi les 858 patientes incluses, seules 59 possèdent des données complètes pour toutes les variables. L'hôpital a anonymisé tous les dossiers avant de publier le dataset. L'ensemble des données est disponible sur le site Web du Machine Learning Repository de l'Université de Californie à Irvine (UCI ML) (University of California Irvine, 1987), avec description des fonctionnalités.

2.2. Precprocessing :

2.2.a. *Exclusion de variables* :

Les variables TimeSinceFirstSTDs (Temps écoulé depuis le premier diagnostic d'IST) et TimeSinceLastSTDs (Temps écoulé depuis le dernier diagnostic d'IST) présentent un taux de données manquantes élevé (91 %). Le manque d'information limite leur valeur prédictive ; elles ont été exclues du dataset. Les autres variables ont toutes moins de 13,64 % de valeurs manquantes. Elles ont été imputées à l'aide de la méthode des équations en chaîne (MICE) (8). Deux variables binaires, STDsCervicalCondylomatosis (antécédent de condylomatosis cervical) et STDsAIDS (antécédent de syndrome d'immunodéficience acquise), ne contiennent que des valeurs uniques ou des données manquantes. Ces variables n'apportent aucune information supplémentaire et ont été supprimées de l'analyse. La variable STDsSyphilis (antécédent de Syphilis) présentant un nombre insuffisant d'événements (moins de 10%), a été exclue de l'analyse multivariée.

La cohorte de patientes contient des duplications, qui ont été conservées. Le nombre limité de variables par rapport au nombre de patientes prévient difficilement les duplications liées au hasard ; les retirer constituerait une perte d'information conséquente.

L'outcome d'intérêt choisi est le diagnostic du cancer cervical. C'est une variable qualitative booléenne nommée DxCancer (TRUE : la patiente a un diagnostic positif au cancer cervical, FALSE : la patiente a un diagnostic négatif au cancer cervical). Le dataset a été diviser en 2 groupes train et test à 80% et 20% respectivement avec un seed fixé à 123 (set.seed(123)).

2.2.b. *Transformations de variables* :

Pour capturer l'influence des variables qualitatives, une transformation (dummification) a été appliquée. La dummification consiste à remplacer les valeurs "True" et "False" par 1 et 0 respectivement, pour indiquer la présence ou l'absence de l'événement de la variable (9). La totalité des variables qualitatives binaires du dataset ont été encodées (variable dummy codée 0 = évènement absent, variable dummy codée 1 = évènement présent).

Les variables STDsVaginalCondylomatosis (antécédent de condylomatosis vaginal), STDsVulvoPerinealCondylomatosis (antécédent de condylomatosis périnéal) et STDsCondylomatosis (antécédent de condylomatosis) présentent une forte colinéarité (corrélation > 80 %). Seule STDsCondylomatosis a été conservée, car elle englobe les informations des trois variables. La variable binaire PelvicInflammatoryDisease (Maladie inflammatoire pelvienne) a également montré une forte corrélation (> 80 %) avec la totalité des variables infectieuses. Elle a été retirée de l'étude.

Les variables infectieuses (STDsMolluscumContagiosum, STDsGenitalHerpes, STDsHPV, STDsHepatitisB et STDsHIV) sont fortement corrélées (> 80 %). Une agrégation permet de conserver une vision globale des infections virales tout en réduisant la complexité des données. La variable ViralInf_HPV_HBV_GenHerp_Molluscum a été créée. Pour chaque valeur de 1 (= présence de l'événement) dans une des variables infectieuses, la variable nouvellement créée se voit assigner 1 (= présence de l'événement). Elle regroupe les informations des variables infectieuses.

Pour minimiser les problèmes de comportement de l'algorithme liés aux différentes plages de valeurs de chaque variables, une mise à l'échelle de toutes les données a été faite en utilisant la normalisation (Z-score).

2.3. Statistical analysis :

Des statistiques descriptives ont été utilisées pour estimer des paramètres tels que les paramètres de dispersion et de tendance centrale pour les variables quantitatives ainsi que les fréquences absolues et relatives pour les variables qualitatives. L'imputation multiple a été réalisée avec le package MICE dans le logiciel R (10). 5 versions complètes des données sont générées après l'imputation (m = 5), le seed de l'imputation à été fixé a 10 (seed = 10) et 5 itérations de l'algorithme d'imputation (maxit = 5). le test T à été utilisé pour comparer les groupes de patientes avant et après imputation. Les corrélations entre variables ont été analysées par un diagnostic de colinéarité effectué grâce au coefficient de Pearson.

Afin d'évaluer la normalité des variables continues, le test de Shapiro-Wilk a été utilisé et les distributions ont été visualisées à l'aide de graphiques QQ-plot. Les hypothèses d'homogénéité des variances ont été évaluées à l'aide des tests de Levene et de Fligner-Killeen. le test de Wilcoxon a été utilisé pour comparer le statut de diagnostic du cancer entre les variables quantitatives. Pour quantifier l'ampleur des différences observées, le coefficient d de Cohen a été calculé pour chaque variable continue. L'association entre les variables qualitatives et le statut de diagnostic du cancer a été évaluée à l'aide du test du Chi-carré de Pearson. L'effet de taille pour ces associations a été mesuré à l'aide du test de V de Cramer. Les rapports de côtes univariés (odds ratios) ont été calculés à l'aide de modèles de régression logistique univariée.

La dummification et la normalisation (Z-core) ont été appliquées pour améliorer les conditions d'application de l'analyse multivariée. La normalisation à été réalisée avec la fonction scale() du package heatmaply (11). Les variables avec une valeur p inférieure à 0,1 dans l'analyse univariée ont été sélectionnées pour être incluses dans l'analyse multivariée. Pour évaluer les principales sources de variation des données, une analyse en composantes principales (PCA) a été réalisée pour explorer les relations entre les variables quantitatives et réduire leur dimensionnalité. Le statut de DxCancer a été intégré en tant que variable qualitative supplémentaire pour évaluer son influence dans l'espace factoriel. Une analyse des correspondances multiples (MCA) a aussi été réalisée pour examiner les relations entre les variables qualitatives. Enfin, une analyse factorielle des données mixtes (FAMD) a finalement été menée, afin d'intégrer simultanément les variables quantitatives et qualitatives, permettant une analyse conjointe.

pour tester l'influence simultanée des variables sur le statut du diagnostic du cancer, des modèles de régression logistique multiple ont été réalisés. Ces modèles multivariés ont permis le calcul des odds ratios avec leur intervalle de confiance à 95%. Les observations étaient indépendantes, et aucune mesure répétée ou série temporelle n'était incluse dans l'analyse. La sélection des variables explicatives a été réalisée par une méthode d'élimination backward basée sur le critère d'information d'Akaike (AIC), puis dans un second temps basée sur le critère (BIC).

La qualité de l'ajustement du modèle a été évaluée par des tests de vraisemblance et des indices de pseudo-R^2, incluant les mesures de McFadden, Cox-Snell, et Nagelkerke. Le test d'adéquation de Hosmer-Lemeshow a

été utilisé pour vérifier la concordance entre les prédictions du modèle et les observations.

le niveau de significativité était de $p < 0.05$ pour toutes les comparaisons. Le r de Pearson a été interprété comme suit : petit effet $r \leq 0,20$, effet moyen $r \geq 0,50$ et effet important $r \geq 0,80$. Les pseudo r de McFadden, Cox-Snell, et Nagelkerke ont été interprété comme suit : petit effet $r \leq 0,10$, effet moyen $r \geq 0,50$ et effet important $r \geq 0,80$. Le d de Cohen a été interprété comme un petit effet $d \leq 0,20$, un effet moyen $d \geq 0,50$ et un effet important $d \geq 0,80$. Le V de Cramer a été interprété comme un petit effet $v \leq 0,20$, un effet moyen $v \geq 0,50$ et un effet important $v \geq 0,80$.

3. Resultats :

3.1. Preprocessing

L'imputation ayant été réalisé en amont des analyses, la Table 1 résume les caractéristiques de la population des patientes avant imputation et après imputation. Nous pouvons y observer que les différences significatives entre les groupes de patientes ont été conservées par l'algorithme d'imputation. Les variables concernées sont le nombre d'années de contraception sous DIU et l'âge du premier rapport (IUDYeras : avant imput. : p-values = 0.002 ; après imput. : p-value < 0.001 , FirstSexualIntercourse : avant imput. : pvalues = 0.002 ; après imput. : pvalue = 0.002). A la lecture de la Table 2, on peut constater que les données imputées ont une distribution significativement différente des données originales pour la variable Hormonal contraceptive years (pvalue = 0.004). Le detail de la distribution avant et apres imputation de cette variable est montrée dans la Figure 1.

3.2. Caractéristiques patients

Cette étude rétrospective inclue 858 patientes. L'âge médian était de 25 ans, avec un intervalle interquartile (IQR) de 20 à 32 ans, tandis que la moyenne d'âge était de 27 ans, avec une plage allant de 13 à 84 ans. Le nombre moyen de grossesses par femme était de 2,19, avec un nombre d'enfant allant de 0 à 11. Le nombre médian de grossesses était de 2 (IQR : 1 à 3). Concernant les antécédents d'infections sexuellement transmissibles (IST), 82 % des patientes (n = 706) n'ont rapporté aucun épisode d'IST. Le diagnostic d'infection par le papillomavirus humain (HPV) a été posé chez 2,1 % des patientes (n = 18). En ce qui concerne le tabagisme, la majorité des patientes (86 %, n = 726) ont fumé pendant une période allant jusqu'à un an. L'utilisation de contraceptifs hormonaux était rapportée par 56 % des patientes (n = 481/858), dont 43 % (n = 367/858) les ont utilisés pendant une période de 0 à 5 ans et 10 % (n = 87/858) de 5 à 10 ans. L'utilisation d'un dispositif intra-utérin (DIU) était peu fréquente, avec 90 % des patients (n = 775/858) n'ayant jamais utilisé de DIU. Le nombre de partenaires sexuels rapportés était compris entre 1 et 5 chez la majorité des patients (94 %, n = 808/858). Ces données permettent d'établir le profil sociodémographique et comportemental des patientes inclus dans cette étude.

Table 3 : Caractéristiques des patientes

Variable	***N = 858***[1]
Age	25 (20, 32)
SmokesYears	
0 to 1	726 (86%)
1 to 5	51 (6.0%)
5 to 10	30 (3.6%)
10<	38 (4.5%)
Unknown	13
HormonalContraceptivesYears	
0	269 (36%)
0 to 5	367 (49%)
5 to 10	87 (12%)
10<	27 (3.6%)
Unknown	108
IUDYears	
0	658 (89%)
0 to 5	55 (7.4%)
5 to 10	21 (2.8%)
10<	7 (0.9%)
Unknown	117
NumberOfSexualPartners	
1 to 5	808 (97%)
5 <	24 (2.9%)
Unknown	26
NumOfPregnancies	
0	16 (2.0%)
1 to 5	758 (95%)
5 <	28 (3.5%)
Unknown	56

[1] *Median (IQR); n (%)*

Les patientes diagnostiquées avec un cancer cervical (n = 18/858 ; 2%) présentaient un âge significativement plus élevé que les patientes non diagnostiquées (moyenne = 33 ans vs moyenne = 27 ans ; $p < 0,001$). De même, ces patientes avaient un âge au premier rapport sexuel plus avancé (moyenne = 18,28 ans vs moyenne = 16,95 ans ; $p = 0,002$). Par ailleurs, les patientes ayant un nombre d'années de contraception au dispositif intra-utérin (DIU) supérieur étaient moins susceptibles d'être diagnostiquées avec un cancer cervical (moyenne = 1,72 ans vs moyenne = 0,42 ans ; $p < 0,001$).

Les participantes diagnostiquées d'un cancer cervicale étaient également plus susceptibles d'avoir un diagnostic antérieur d'HPV positif (DxHPV : 89 % vs 0,2 % ; $p < 0,001$), un diagnostic antérieur global positif (Dx : 78 % vs 1,2 % ; $p < 0,001$) et des résultats positifs aux tests de dépistage tels que Hinselmann (22 % vs 3,7 % ; $p = 0,005$), Schiller (39 % vs 8 % ; $p < 0,001$), cytologie (22 % vs 4,8 % ; $p = 0,011$) et biopsie (33 % vs 5,8 % ; $p < 0,001$).

En revanche, aucun lien significatif n'a été observé avec les participantes diagnostiquées d'un cancer cervicale concernant le nombre moyen de partenaires sexuels, le nombre de grossesses, ou les antécédents de consommation de tabac (toutes p > 0,05).

3.3. Analyse Multivariée

Une analyse des composantes principale révèlent que les 2 premières composantes, expliquent ensemble 44,6 % de la variation du dataset.

Figure 5 : Analyse factorielle sur données mixtes (variables)

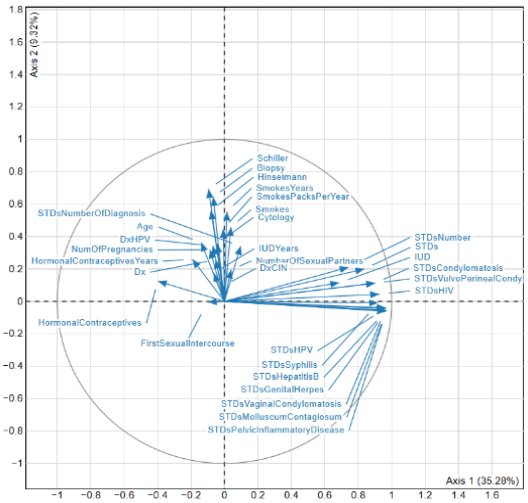
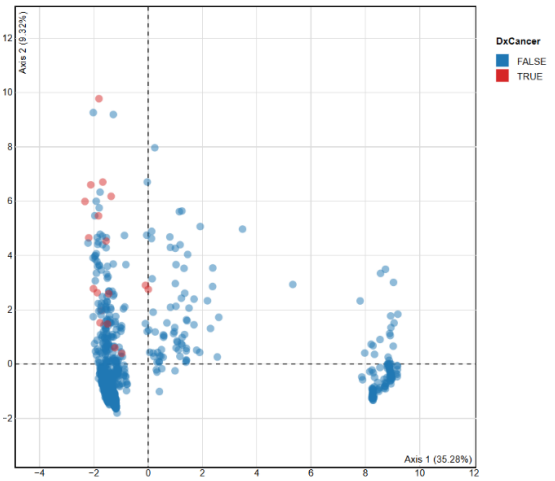


Figure 6 : Analyse factorielle sur données mixtes (individus)



Une analyse de régression logistique a été réalisée pour estimer les rapports de cotes (OR) associés aux différentes caractéristiques cliniques et comportementales, permettant d'évaluer leur relation avec le diagnostic de cancer. Les résultats sont présentés dans la table 4, cependant nous pouvons noter un intervalle de confiance inutilisable.

Table 4 : Présentation des Odds Ratios

Characteristic	**OR[1]**	**95% CI[1]**	**p-value**
Age	0.01	0.00, Inf	>0.9
NumberOfSexualPartners	10.8	0.00, Inf	>0.9
FirstSexualIntercourse	2,674	0.00, Inf	>0.9
NumOfPregnancies	237		>0.9
Smokes	0.00	0.00, Inf	>0.9
SmokesYears	38.9	0.00, Inf	>0.9
SmokesPacksPerYear	2.70	0.00, Inf	>0.9
HormonalContraceptives	inf	0.00, Inf	>0.9
HormonalContraceptivesYears	192	0.00, Inf	>0.9
IUD	0.00	0.00, Inf	>0.9
IUDYears	8,552	0.00, Inf	>0.9
STDs	inf	0.00, Inf	>0.9
STDsNumber	0.00	0.00, Inf	>0.9
STDsCondylomatosis	523,349		>0.9
STDsSyphilis	0.00	0.00, Inf	>0.9
STDsNumberOfDiagnosis	0.00	0.00, Inf	>0.9
DxCIN	0.00	0.00, Inf	>0.9
DxHPV	inf	0.00, Inf	>0.9
Dx	inf	0.00, Inf	>0.9
Hinselmann	0.71	0.00, Inf	>0.9
Schiller	0.00	0.00, Inf	>0.9
Cytology	0.00	0.00, NA	>0.9
Biopsy	inf	0.00, Inf	>0.9
Viral_inf_HPV_HBV_GenHerp_Molluscum	63.7		>0.9

[1] OR = Odds Ratio, CI = Confidence Interval

4. Discussion

Dans cette analyse, nous avons utilisé des outils statistiques pour identifier les facteurs de risque associés au cancer du col de l'utérus à partir d'une base de données hospitalière issue de l'hôpital universitaire de Caracas, au Venezuela. Les résultats mettent en évidence plusieurs variables pertinentes, tout en soulignant de nombreuses limites méthodologiques.

4.1. Points forts de l'étude

L'un des principaux atouts de ce travail est l'utilisation de variables potentiellement reproductibles en pratique clinique. Comme le souligne Hendriksen et al. (2013), un modèle prédictif fiable doit s'appuyer sur des données issues d'une collecte rigoureuse et applicable dans un large éventail de contextes cliniques. Les données fournies par Fernandes et al. (2017) repose sur des variables cliniques et comportementales standardisées, souvent utilisées dans la prise en charge des patientes à risque. De plus, la méthode de détermination de l'outcome a été réalisée en s'appuyant sur des tests diagnostiques validés (Hinselmann, Schiller, cytologie, biopsie), garantissant une bonne précision dans la classification des patientes.

L'utilisation de techniques de traitement des données, telles que l'imputation multiple (MICE) pour gérer les valeurs manquantes, l'analyse en composantes principales (PCA) et l'analyse factorielle des données mixtes (FAMD) pour réduire la dimensionnalité des variables, contribue à la robustesse de cette analyse. De plus, nous avons appliqué une méthodologie statistique rigoureuse, incluant des tests de colinéarité, la normalisation des variables et une sélection des variables basée sur les critères d'information d'Akaike (AIC) et de Bayes (BIC), permettant ainsi une meilleure interprétation des résultats.

4.2. Limites de l'étude

Cependant, cette analyse présente plusieurs limites qui doivent être prises en compte pour l'interprétation et l'applicabilité des résultats. Tout d'abord, la totalité des données a été collectée auprès d'un unique hôpital universitaire à Caracas, ce qui limite la représentativité de notre échantillon. Cette restriction géographique et institutionnelle empêche une généralisation facile des résultats à d'autres populations ayant des caractéristiques socio-économiques, culturelles et médicales différentes. Des études multicentriques, incluant des cohortes issues de différents pays et contextes cliniques, seraient nécessaires pour renforcer l'extrapolation des résultats de cette analyse.

Une autre limitation importante concerne la présence de valeurs manquantes. Plusieurs patientes ont choisi de ne pas répondre à certaines questions pour des raisons de confidentialité. Bien que nous ayons utilisé l'imputation multiple pour réduire l'impact de ces valeurs manquantes, cela pourrait introduire un biais si les données manquantes ne sont pas totalement aléatoires. De plus, nous avons pu constater que les données imputées ont une distribution significativement différente des données originales pour la variable Hormonal contraceptive years (p-value = 0.004) ce qui induit un biais lors du traitement ce ces données. Pour vérifier si l'imputation influence les conclusions, les prochaines analyses devront inclure une analyse de sensibilité qui consiste à comparer les résultats d'analyses effectuées sur le dataset imputé avec ceux effectués sur le dataset non imputé.

Certaines variables ont dû être exclues en raison d'un taux trop élevé de données manquantes (ex. : TimeSinceFirstSTDs et TimeSinceLastSTDs), ce qui a pu entraîner une perte d'information pertinente.

Un autre défi concerne l'utilisation d'un modèle de régression logistique classique, qui repose sur plusieurs hypothèses, notamment l'absence de forte colinéarité entre les variables explicatives et une relation linéaire entre les prédicteurs et le logarithme des odds. Nos analyses ont révélé une forte colinéarité entre certaines variables, nécessitant des transformations et agrégations (ex. : regroupement des infections virales dans une variable unique).

4.3. Perspectives et recommandations

Cette analyse souligne l'importance d'intégrer des données de qualité pour construire et vérifier des hypothèses afin d'améliorer le dépistage du cancer du col de l'utérus. À l'avenir, il serait pertinent de comparer ces résultats à d'autres études statistiques incluant d'autres outils, telles que les réseaux neuronaux ou les méthodes d'apprentissage profond. Ces techniques ont démontré leur efficacité dans l'identification de variables pouvant aider à la prédiction de diagnostics médicaux, à partir de données cliniques et biologiques.

En conclusion, notre étude met en évidence plusieurs facteurs de risque associés au cancer cervical comme l'âge ou un antécédant d'infection à HPV. Toutefois, la généralisation des résultats doit être faite avec précaution en raison des limites liées à la représentativité de l'échantillon et aux valeurs manquantes. Des études complémentaires, idéalement multicentriques et intégrant des méthodes de modélisation plus avancées, sont nécessaires pour affiner la prédiction du risque et optimiser les stratégies de prévention et de dépistage.

Bibliography :

1. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011 Oct;18(10):1099–104.
2. Harris JK. Statistics With R: Solving Problems Using Real-World Data. SAGE Publications; 2019. 821 p.
3. Singh D, Vignat J, Lorenzoni V, Eslahi M, Ginsburg O, Lauby-Secretan B, et al. Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative. Lancet Glob Health. 2022 Dec 14;11(2):e197–206.
4. risk_factors_cervical_cancer.CSV
5. Fernandes, Cardoso & Fernandes (2017b).Fernandes K, Cardoso JS, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening. In: Alexandre L, Salvador Sánchez J, Rodrigues J, editors. Iberian Conference on Pattern Recognition and Image Analysis; Faro: Springer; 2017b. pp. 243–250.
6. Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative
7. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. ResearchGate. 2024 Oct 22
8. White, I. R., P. Royston, and A. M. Wood. 2011. “Multiple Imputation Using Chained Equations: Issues and Guidance for Practice.” Statistics in Medicine 30 (4): 377–99.
9. Ernsten L, Körner LM, Heil M, Schaal NK. The association between 2D:4D digit ratio and sex-typed play in children with and without siblings. Sci Rep. 2024 Jul 2;14:15231.
10. Package R : MICE
11. Package R : heatmaply
12. Hendriksen JM, Geersing G, Moons K, et al. Diagnostic and prognostic prediction models. J Thromb Haemost 2013;11:129-41.

Supplementary :

Table 1 : Caractéristiques avant/après imputation.

	Before Imputation			After Imputation		
Variable	**DxCancer***FALSE* (n=840)[1]	**DxCancer***TRUE* (n=18)[1]	**p-value**[2]	**DxCancer***FALSE* (n=840)[1]	**DxCancer***TRUE* (n=18)[1]	**p-value**[2]
Age	25 (20, 32)	32 (28, 38)	<0.001	25 (20, 32)	32 (28, 38)	<0.001
SmokesYears	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.9	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.8
Unknown	12	1				
SmokesPacksPerYear	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.8	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.7
Unknown	12	1				
HormonalContraceptivesYears	0.50 (0.00, 3.00)	1.00 (0.02, 5.25)	0.2	0.25 (0.00, 2.00)	1.00 (0.02, 5.25)	0.083
Unknown	108	0				
IUDYears	0.00 (0.00, 0.00)	0.00 (0.00, 3.00)	0.002	0.00 (0.00, 0.00)	0.00 (0.00, 3.00)	<0.001
Unknown	117	0				
NumberOfSexualPartners	2.00 (2.00, 3.00)	3.00 (2.00, 3.00)	0.3	2.00 (2.00, 3.00)	3.00 (2.00, 3.00)	0.4
Unknown	26	0				
FirstSexualIntercourse	17.00 (15.00, 18.00)	18.50 (18.00, 19.00)	0.002	17.00 (15.00, 18.00)	18.50 (18.00, 19.00)	0.002
Unknown	7	0				
NumOfPregnancies	2.00 (1.00, 3.00)	2.00 (2.00, 3.75)	0.2	2.00 (1.00, 3.00)	2.00 (2.00, 3.75)	0.10
Unknown	56	0				
STDsNumberOfDiagnosis			>0.9			>0.9
0	770 (92%)	17 (94%)		770 (92%)	17 (94%)	
1	67 (8.0%)	1 (5.6%)		67 (8.0%)	1 (5.6%)	
2	2 (0.2%)	0 (0%)		2 (0.2%)	0 (0%)	
3	1 (0.1%)	0 (0%)		1 (0.1%)	0 (0%)	

[1] Median (IQR); n (%)

[2] Wilcoxon rank sum test; Fisher's exact test

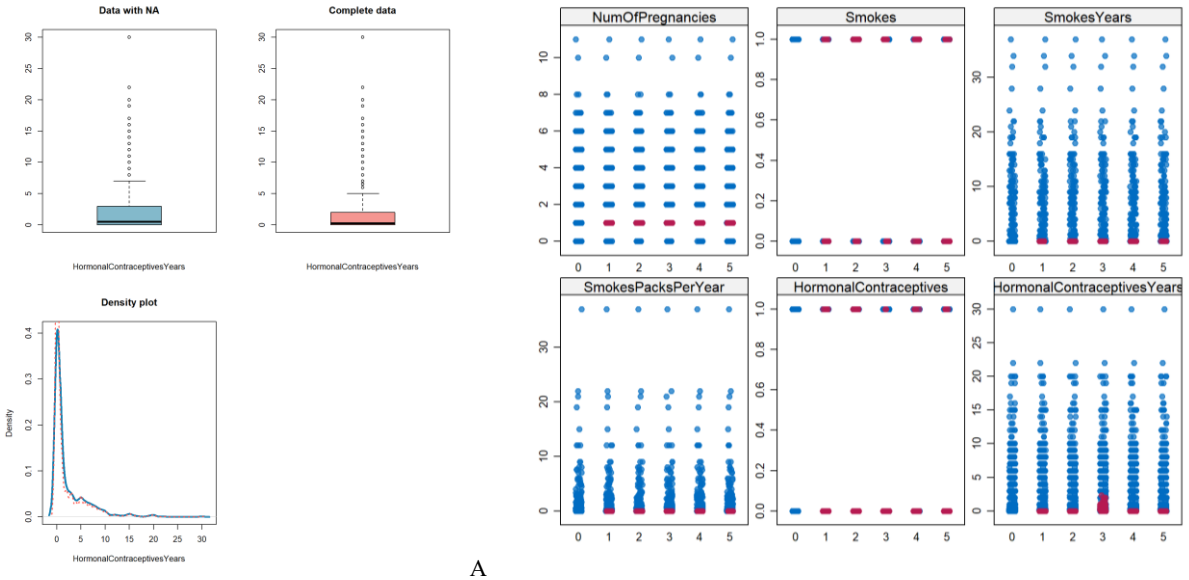
Table 2 : Test de comparaison avant/après imputation.

Variable	**Before Imputation**(n=858)[1]	**After Imputation**(n=858)[1]	**p-value**[2]
Age	25 (20, 32)	25 (20, 32)	>0.9
SmokesYears	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.9
Unknown	13	0	
SmokesPacksPerYear	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.9
Unknown	13	0	
HormonalContraceptivesYears	0.50 (0.00, 3.00)	0.25 (0.00, 2.00)	0.004
Unknown	108	0	
IUDYears	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.3
Unknown	117	0	
NumberOfSexualPartners	2.00 (2.00, 3.00)	2.00 (2.00, 3.00)	0.4
Unknown	26	0	
FirstSexualIntercourse	17.00 (15.00, 18.00)	17.00 (15.00, 18.00)	0.9
Unknown	7	0	
NumOfPregnancies	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	0.14
Unknown	56	0	
STDsNumberOfDiagnosis			>0.9
0	787 (92%)	787 (92%)	
1	68 (7.9%)	68 (7.9%)	
2	2 (0.2%)	2 (0.2%)	
3	1 (0.1%)	1 (0.1%)	

[1] Median (IQR); n (%)

[2] Wilcoxon rank sum test; Fisher's exact test

Figure 1 : Distribution de la variable HormonalContraceptivesYears avant/après imputation.

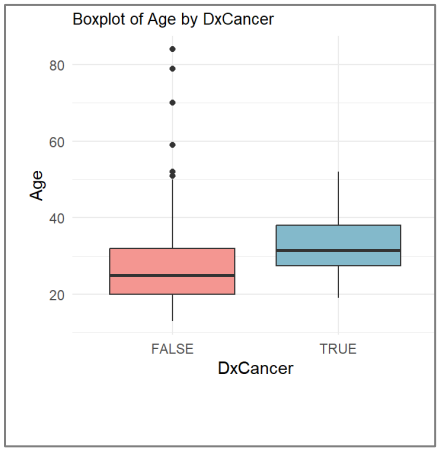


A B

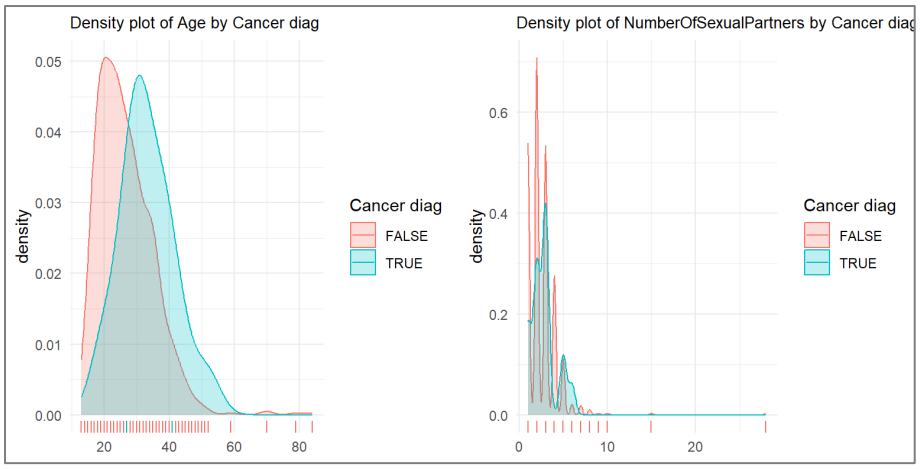
(A). La courbe bleue représente la distribution avant imputation. La courbe rouge représente la distribution après imputation.

(B). les points bleus représentent la distribution avant imputation pour chacun des jeux de données. les points rouges représentent les points imputés générés pour chacun des jeux de données.

Figure 2 : Distribution de l'Age pour les patientes avec et sans diagnostic de cancer cervical.



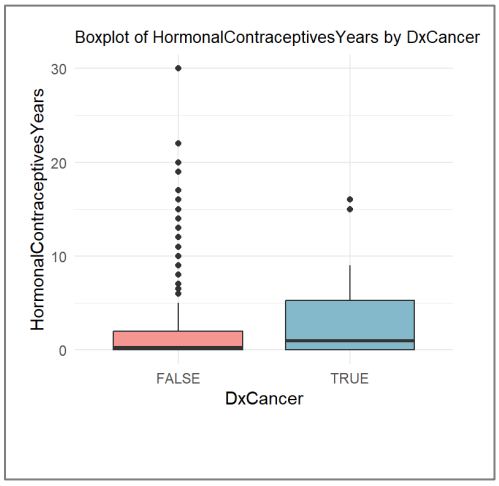
A



B

Figure 3 : Distribution en années de l'utilisation de contraceptifs hormonaux pour les patientes avec et sans diagnostic de cancer cervical.

A



B

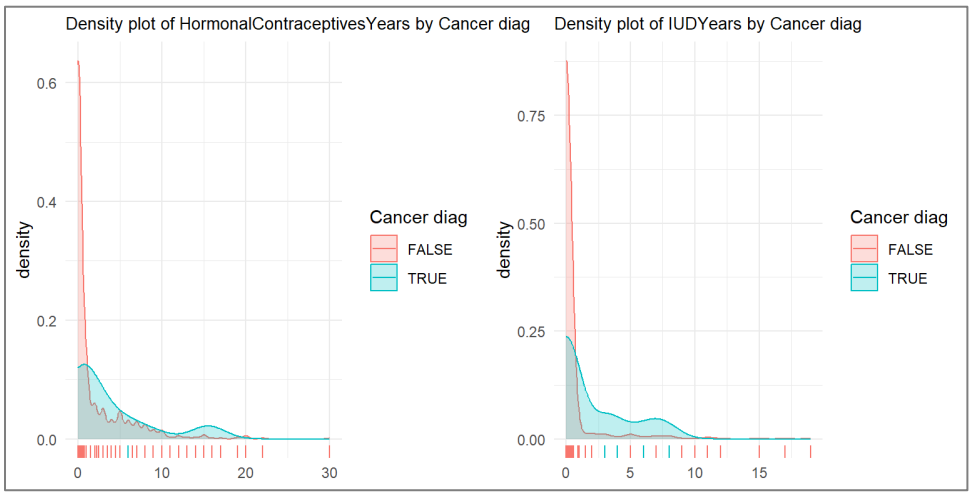


Figure 4 : Distribution des diagnostics au papillomavirus pour les patientes avec et sans diagnostic de cancer cervical.