Análisis Estadístico de la Relación entre Genotipos y Lesiones del Papiloma Humano en Mujeres de la Región Central de México

Rebeca K. Torres-Septien[1] and B. Itzelt Gómez-Catzín[1]

[1] *Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias*

Reception date of the manuscript: 1/11/2023
Acceptance date of the manuscript: 24/11/2023
Publication date: November 25, 2023

Abstract—Despite the widespread impact of human papillomavirus (HPV) and its established connection to cervical cancer (CC), particularly in Central Mexico, the research remains scarce. Our study focuses on the spread of different HPV types in Central Mexico and how they relate to the development of cervical cancer. We analyzed data from 883 women, looking at their age, the presence of HPV genotypes, and the severity of cervical changes. Our goal was to see if older age was linked to more serious cervical conditions and to identify any patterns. Through our analysis, we aim to contribute valuable insights into the age-related progression of HPV-linked dysplasia and offer predictive modeling tools to aid in the early detection and tailored treatment of this disease. The findings advocate for a more robust diffusion of sexual health education and treatment approaches to address the most prevalent and harmful HPV genotypes in Mexico.

Keywords— Cáncer Cervical, Papiloma Humano, Genotipos, Correlaciones, Mujeres Mexicanas, Análisis Estadístico.

I. Introducción

El Cáncer Cervical (CC) es una de las principales causas de muerte en las mujeres que viven en países en desarrollo. [1] Se estima que cada año ocurren 500,000 casos nuevos a nivel mundial, y aproximadamente 270,000 mujeres mueren debido a esta enfermedad; muchos de estos sucesos ocurren en América Latina, siendo México el segundo lugar con cáncer cervical. En 2008, se aproxima que 10,186 mujeres mexicanas desarrollaron la enfermedad y como resultado, 5061 fallecieron. [2]

Asimismo, la abundante literatura a nivel global registra que los genotipos 16 y 18 son los más importantes en prevalencia y potencial ontogénico, no obstante, numerosos artículos han reportado que en México los genotipos 16, 31, y 51 son más predominantes, considerando que el 33 y 52 también fueron más comunes, mientras que el 18 tuvo menos prevalencia, lo cual nos indica que existe un gran índice de genotipos no incluidos en las vacunas aplicadas actualmente, esto si consideramos que en el año 2011 el Consejo Nacional de Inmunización aprobó la expansión de las campañas de vacunación de VPH a escuelas que incluían a todas las niñas de 9 y 11 años, este programa usaba vacunas contra VPH 16/18 y VPH 6/11/16/18. [3]

Esta cobertura ha ido incrementando a lo largo del tiempo, ya que según la información reportada en 2018, 1 millón de dosis fueron aplicadas en todo el país, sin embargo, esta cifra es muy baja si consideramos a toda la población de 125 millones de personas, cuyo 5.7% son mujeres de entre 9 y 14 años de edad. [4]

Igualmente, estudios en diferentes regiones han mostrado una mayor prevalencia de VPH en universitarias y mujeres menores a 25 años debido a una conducta sexual arriesgada, falta de conocimiento de infecciones por VPH, baja tasa de vacunación, entre otros factores socioculturales. De hecho, existen estudios donde a pesar de estar casi en la misma zona geográfica, tienen resultados con ciertas diferencias; con ello se pudo determinar que dependiendo de la región geográfica algunos genotipos son más frecuentes que otros, también, las diferentes características sociodemográficas de las participantes pueden explicar las disimilitudes en los efectos de vacunación de ambas poblaciones. [4]

Este estudio fue diseñado en respuesta a la creciente necesidad de información para la investigación continua, análisis y prevención de la infección por el Virus del Papiloma Humano. Se centra en identificar y comprender la prevalencia, distribución y detección de las variables que influyen en el desarrollo de la enfermedad. El objetivo principal es proporcionar un análisis estadístico que sirva como material complementario para investigaciones actuales y futuras, buscando identificar patrones entre las variables para desarrollar estrategias de prevención y detección de esta enfermedad.

Estos esfuerzos subrayan la vital importancia del análisis, prevención y difusión de información acerca de esta enfermedad. Es crucial considerar que los grupos más vulnerables de mujeres en México están siendo afectados desde edades tempranas debido a la falta de educación sexual e investigaciones que respalden la implementación de nuevas vacunas

y tratamientos contra los genotipos más predominantes en el país. Esta conexión entre los resultados de investigaciones pasadas, el diseño de nuestro estudio y las necesidades identificadas, refuerza la relevancia de abordar de manera integral la problemática del VPH en la población mexicana.

La investigación seleccionó una base de datos que reporta a 883 casos de mujeres de distintas edades con diferentes genotipos de VPH con bajo o alto riesgo en el centro de México. Para ser más precisos, el estudio fue desarrollado en el estado de Aguascalientes, que colinda con las ciudades de los estados de Jalisco y Zacatecas; se hizo un gran énfasis de ello, por lo cual podemos atribuir que las mujeres de dichos estados pudieron haber asistido a la Clínica Displásica del Hospital General de Zona Número 1. [2].

La base de datos presentaba la edad de cada mujer junto a el tipo de genotipo clasificado de forma binaria(0 si no presentaba el genotipo y 1 si sí) el tipo de displasia y sus clasificaciones: Cancer Cervical Agresivo (ICC), Lesión Intraepitelial Escamosa de Alto Grado (HGSIL), Lesion Escamosa Intraepitelial de Bajo Grado (LGSIL), y Negativo para Lesión Intraepitelial (NILM).

En función de ello, buscamos explorar si hay una relación entre las diferentes lesiones causados por el VPH y los genotipos, además de relacionar esto con la edad de cada mujer, es decir, se busca encontrar la relación entre las mujeres de edad más avanzada y el tipo de displasia. Adicionalmente, investigamos si existe una tendencia de ciertos genotipos de VPH a agruparse con tipos específicos de lesiones.

Con el objetivo de identificar las tendencias en nuestros datos, implementaremos conceptos y métodos estadísticos para analizar las variables contenidas en nuestra base de datos. Este análisis se efectuará a través de una serie de correlaciones y asociaciones, con la finalidad de desarrollar un modelo predictivo. Dicho modelo estará diseñado para estimar, basándose en la edad y el genotipo de la paciente, el tipo de displasia que podría desarrollarse. Este enfoque tiene como último propósito facilitar la aplicación de tratamientos adecuados y prevenir la progresión del daño ocasionado por la infección del VPH.

II. Metodología

a. Análisis de Datos Estadísticos

Atrevas del lenguaje de programación Python, utilizamos las librerías de *pandas*, *matplotlin* y *seaborn* para encontrar la frecuencia de los genotipos en la base de datos, la correlación entre nuestras variables de displasia y genotipos, así como la relación entre la frecuencia y la correlación con el tipo de displasia, ello con la finalidad de entender el comportamiento de cada variable en nuestra base de datos. Igualmente, al trabajar con valores numéricos, se implementó una codificación de la variable Tipo de Displasia (Tabla 1), es decir, asignamos un valor numérico, del 0 a 3, dependiendo de qué tan agresivo fuera la lesión.

Como se mencionó anteriormente, nuestra base de datos está constituida por 883 mujeres con diferentes edades, siendo el mínimo de 15 y máximo de 72 y la edad promedio de 35 años. Para comprender mejor la distribución de los genotipos y el tipo de displasia en el estudio, presentamos los gráficos de barras correspondientes (Figuras 1 y 2) donde se destaca la presencia de los genotipos 16, 31 y 51, lo cual concuerda con los hallazgos en los artículos estudiados. Por otro lado, sobresale la abundancia de displasia NILM y LGSIL con una proporción de 349 y 248 para cada caso.

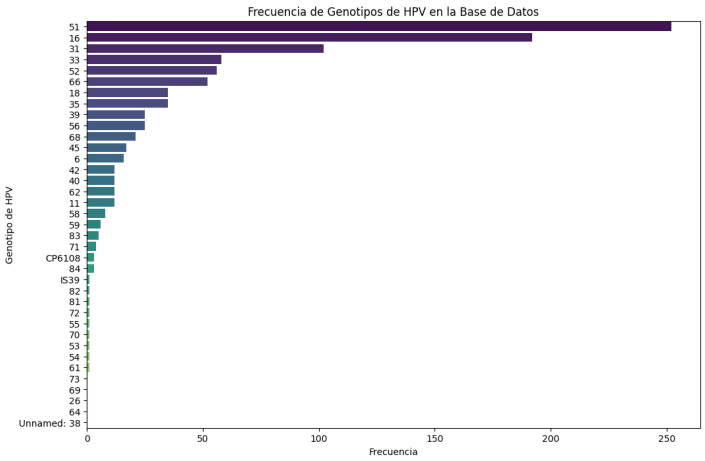

Fig. 1: Frecuencia de Genotipos en la Base de Datos

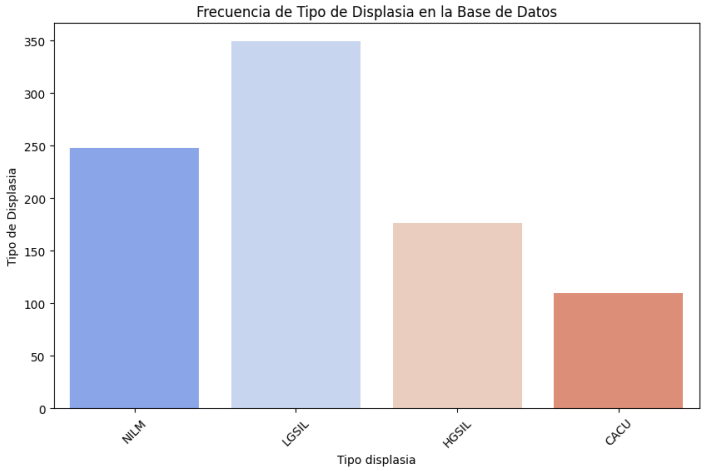

Fig. 2: Frecuencia del Tipo de Displasia en la Base de Datos

b. Correlación entre Genotipos y el Tipo de displasia

Primero, para entender el daño que el virus tiene sobre la población femenina, necesitamos encontrar si existe una tendencia entre los genotipos y el tipo de displasia, esto con la finalidad de determinar si existen ciertos genotipos que se manifiestan de forma más agresiva. Se desarrollaron dos matrices para cada variable para calcular la correlación entre ambas, como podemos ver en la Figura 3, interpretamos lo siguiente

Table 1: Codificación del Tipo de Lesión

Tipo de Lesión	Valor numérico	Daño
NILM	0	Nulo
LGSIL	1	Leve
HGSIL	2	Severo
ICC	3	Cancerígeno
CACU	3	Cancerígeno

- Genotipos con Mayor Correlación Positiva: Los genotipos 51, 16 y 33 tienen las correlaciones más altas y positiva con el tipo de displasia. Esto indica que estos están más fuertemente asociados con tipos de displasia más severos (HGSIL, ICC, CACU)
- Genotipos con Correlación Negativa o Baja: Algunos genotipos, como el 6 y 11 muestran una correlación negativa muy baja, lo que sugiere una asociasión menos significativa, incluso inversa con los tipos de displasia más severos.
- Genotipos sin Correlación Significativa: Los genotipos 26, 64, 69 y 73 no tienen correlaciones calculadas debido en gran parte a la falta de variabilidad o presencia en la muestra.

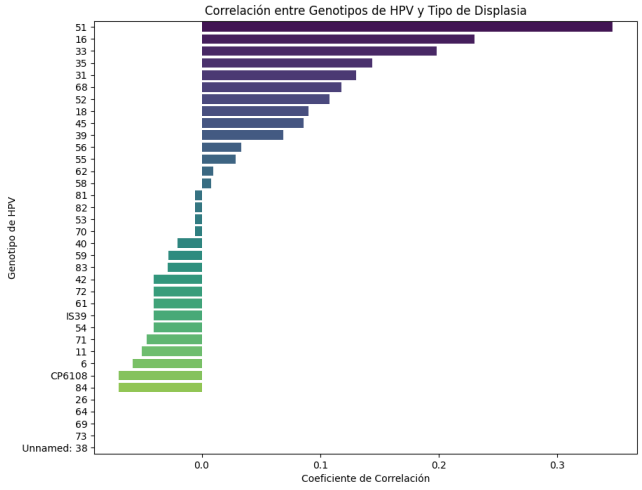

Fig. 3: Correlación entre Genotipos de VPH y Tipo de Displasia

Un genotipo con una correlación positiva alta no necesariamente significa que es un causante de las formas más severas de displasia, pero podría ser un factor de riesgo o un marcador biológico asociado con la severidad de la enfermedad. La variedad en las magnitudes de correlación sugiere que diferentes genotipos del VPH pueden tener distintos niveles de influencia en la progresión de las lesiones cervicales y el cáncer cervical. Los genotipos con fuertes correlaciones positivas podrían ser objetivos claves en el diagnóstico temprano y la prevención del cáncer cervical. Su identificación en pruebas de VPH podría ayudar a categorizar a los pacientes en grupos de riesgo más alto para una vigilancia más estrecha.

c. Correlación entre Edad y Tipo de Displasia

Al calcular el coeficiente de correlación entre la edad de las mujeres y el tipo de displasia, obtuvimos un valor de 0.4580, en un rango de 0 a 1, donde 0 indica ninguna correlación y 1 indica una correlación perfecta. Este resultado implica una correlación moderada entre la edad y la severidad de la displasia. Para visualizar mejor esta relación, representamos los datos con un *Box-Plot* (ver Figura 4), que ilustra que a medida que aumenta la edad, la severidad de la displasia tiende a incrementarse. Podemos ver que mujeres entre los 40 y 50 años tienen una mayor probabilidad de desarrollar una displasia cancerígena.

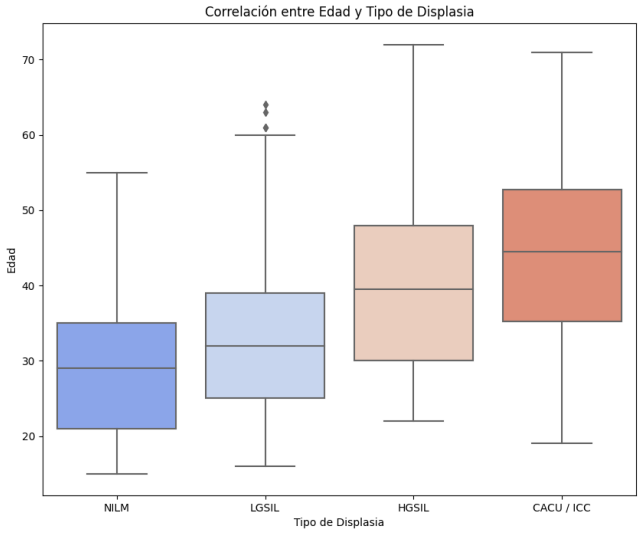

Fig. 4: Correlación entre Genotipos de VPH y Tipo de Displasia

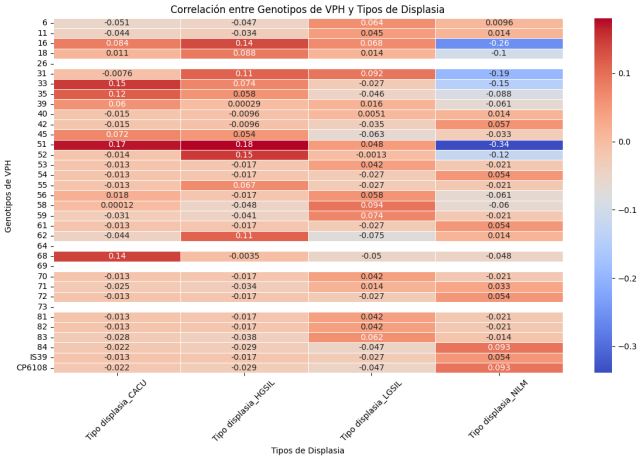

Fig. 5: Correlación entre Genotipos de VPH y Tipos de Displasia

d. Modelo con Regresión Logística

El conjunto de datos abarca múltiples variables, pero para este análisis especifico , se centró en la edad como la variable independiente principal y el tipo de displasia cervical codificado(NILM, LGSIL, HGSIL, ICC, CACU) como la variable dependiente. La metodología empleada involucró la aplicación de un modelo de regresión logística multinomial (Ver 2), facilitada por el uso de la biblioteca *statsmodel* en Python. Esta técnica es particularmente adecuada para datos categóricos y permite modelar la probabilidad de cada categoría de displasia en función de la edad (Ver Figura 6).

1. Resultados clave

El modelo reveló que la edad es un predictor significativo del Cáncer Cervical. Cada categoría de displasia mostró una relación positiva con la edad, indicando que a medida que aumenta la edad, también lo hace la probabilidad logarítmica de tener una displasia mas grave en comparación con la categoría de referencia (NILM). Los resultados estadísticamente significativos, dados los valores de p bajos asociados con los coeficientes de edad en cada categoría, refuerzan la relevancia de edad como factor en la incidencia de displasia cervical.

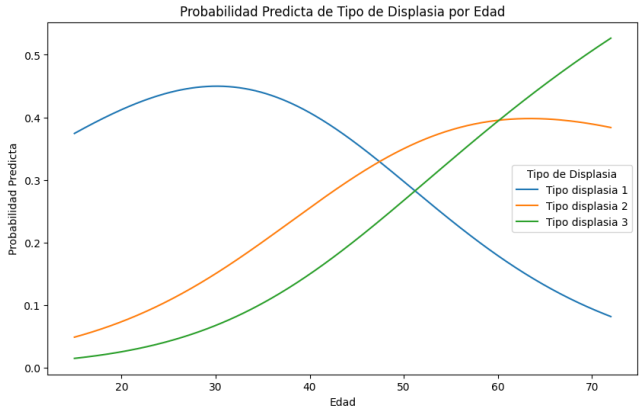

Fig. 6: Probabilidad Predicta de Tipo de Displasia por Edad

2. Interpretación

La interpretación de los coeficientes sugiere que existe una correlación positiva entre la edad y el riesgo de presentar formas más avanzadas de displasia cervical. En términos prácticos, esto implica que las mujeres mayores tienen una probabilidad incrementada de presentar tipos de displasia más severos. Esto es coherente con la comprensión médica de que el riesgo de cáncer cervical y sus lesiones precursoras aumenta con la edad.

e. Modelo con predicción

Desarrollamos un modelo con el cual sirve para predecir el tipo de displasia tomando como variables independientes los genotipos y la edad de la paciente. Esta es una herramienta matemática sumamente útil y aplicable en la área médica para poder determinar el tipo de tratamiento o enfoque clínico a tratar la infección. Dado que nuestra base de datos es relativamente pequeña, la precisión del modelo es de un 60%, lo cual no es óptimo para implementarlo, no obstante este modelo desarrollará una mejor predicción al utilizar cada vez más datos y variables. Como se muestra en 1 , el modelo de predicción utiliza un mapeo de tipos de displasia y luego procede con la división de los datos y el ajuste del modelo.

III. Limitaciones y Consideraciones Futuras

En este artículo, abordamos el análisis de una base de datos que, a pesar de su extensión, carecía de información detallada sobre componentes de la enfermedad. Es crucial destacar que la ausencia de datos personales de las mujeres participantes en el estudio limitaba nuestra capacidad para evaluar el impacto del sistema inmunológico, el cual desempeña un papel fundamental en el desarrollo de la enfermedad. La inclusión de esta variable es esencial, ya que algunos individuos pueden prevenir o eliminar la infección mediante su respuesta inmunológica. No obstante, al considerar esta nueva variable, se hace necesario implementar enfoques más avanzados, como modelos de ecuaciones diferenciales integro-parciales. Estos modelos permitirían comprender la dinámica celular entre las células inmunes principales y las células infectadas, con el objetivo de identificar los factores que favorecen la eliminación de la infección. [5]

Asimismo, el modelo establecido tiene limitaciones inherentes. El Pseudo R-cuadrado, obtenido en el modelo de regresión logística fue relativamente bajo, lo que sugiere que hay otros factores, posiblemente incluyendo distintos genotipos del VPH y factores de riesgo y de comportamiento, que también pueden desempeñar un papel crucial en la determinación de los riesgos de displasia cervical. Por lo tanto, futuros análisis podrían beneficiarse de la inclusión de estas variables adicionales para obtener un modelo más robusto y explicativo.

IV. Resultados

Resaltamos que la edad y los genotipos son factores cruciales en la predicción del riesgo asociado con la displasia cervical con potencial cancerígeno en mujeres mayores, en particular entre los 40 y 50 años. Este hallazgo se alinea con los posibles escenarios asociados con el transcurso del tiempo, incluyendo el cuidado de la salud, la cantidad de parejas sexuales y los hábitos individuales. En este grupo demográfico, observamos patrones que sugieren una interacción compleja entre la edad y factores de estilo de vida, lo cual respalda la importancia de considerar estos elementos al evaluar el riesgo de infección por VPH y el desarrollo de displasia cervical. Además, enfatizamos la elevada presencia de la infección por el genotipo 51, el cual no forma parte del nuevo esquema de vacunación en México. [3] Es importante destacar que, al analizar la relación entre los genotipos, el tipo de displasia y la edad, hemos identificado objetivos clave para el diagnóstico temprano del VPH.

A partir de los análisis presentamos el modelo de regresión logística, el cual demostró una alta correlación entre el tipo de displasia y la edad, como mencionamos anteriormente, esto subraya la necesidad de enfocarse en el tratamiento y la detección temprana del VPH.

V. Conclusiones

La aplicación de estas estadísticas en entornos clínicos podría mejorar notablemente la gestión de citas médicas y los tratamientos del VPH, contribuyendo a una reducción significativa en la tasa de cáncer cervical. No obstante, es crucial destacar que nuestro estudio tiene las limitaciones mencionadas previamente, sugiriendo áreas para futuras investigaciones y mejoras en la eficacia del algoritmo al implementar estas nuevas variables. El objetivo principal de este estudio es evidenciar que la manifestación temprana del Virus del Papiloma Humano, cuando se detecta y trata oportunamente, es menos severa en comparación con los daños que se acumulan a lo largo de los años sin tratamiento.

VI. Agradecimientos

Agradecemos a la Dr. Miriam I. Jiménez Pérez por asesorarnos y encaminarnos durante la investigación médica de este estudio, además de la Doctora Luz M. González Ureña por el asesoramiento en la interpretación de los análisis. Por último, pero no menos importante, al Dr. Edgar René López-Mena por guiarnos y apoyarnos en este estudio.

A. Material Suplementario

El codigo completo y la base de datos se encuentran en este repositorio de Git-Hub: **Repositorio.**

a. Algoritmos

Listing 1: Modelo de predicción

```

displasia_mapping = {'NILM': 0, 'LGSIL': 1, 'HGSIL': 2, 'ICC': 3, 'CACU': 3}
hpv_data['Tipo displasia Encoded'] = hpv_data['Tipo displasia'].map(displasia_mapping)
hpv_data = hpv_data.drop(['Tipo displasia', 'Unnamed: 38'], axis=1)

# Dividir los datos en caracteristicas y objonetivo
X = hpv_data.drop('Tipo displasia Encoded', axis=1).apply(pd.to_numeric, errors='coerce').fillna(0)
y = hpv_data['Tipo displasia Encoded']

# Dividir los datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Ajuste del modelo con busqueda de cuadricula
model = DecisionTreeClassifier()
param_grid = {
    'max_depth': [10, 20, 30, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
grid_search = GridSearchCV(model, param_grid, cv=5, n_jobs=-1, scoring='accuracy')
grid_search.fit(X_train, y_train)

# Evaluar con validacion cruzada
best_model = grid_search.best_estimator_
scores = cross_val_score(best_model, X, y, cv=5, scoring='accuracy')
```

Listing 2: Modelo de regresión

```
import pandas as pd
import statsmodels.api as sm

X = hpv_data[['Edad']]
y = hpv_data['Tipo displasia Encoded']

X = sm.add_constant(X)
model = sm.MNLogit(y, X).fit()
print(model.summary())
```

References

[1] R. Peralta-Rodríguez, P. Romero-Morelos, and V. e. a. Villegas-Ruíz, "Prevalence of human papillomavirus in the cervical epithelium of Mexican women: meta-analysis." *National Library of Medicine*, vol. 7, no. 34, 2012. [Online]. Available: https://doi.org/10.1186/1750-9378-7-34

[2] R. G. Campos, A. Malacara Rosas, E. Gutiérrez Santillán, M. Delgado Gutiérrez, R. E. Torres Orozco, E. D. García Martínez, L. F. Torres Bernal, and A. Rosas Cabral, "Unusual prevalence of high-risk genotypes of human papillomavirus in a group of women with neoplastic lesions and cervical cancer from central mexico," *PLOS ONE*, vol. 14, no. 4, pp. 1–13, 04 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0215222

[3] A. Campos-Romero, K. S. Anderson, A. Longatto-Filho, M. Luna-Ruiz Esparza, D. J. Morán-Porterla, J. A. Castro-Menéndez, J. L. Moreno-Camacho, D. Y. Calva-Espinoza, M. A. Acosta-Alfaro, F. A. Meynard-Mejía, M. Muñóz-Gaitán, and J. Alcántar-Fernández, "The burden of 14 hr-HPV genotypes in women attending routine cervical cancer screening in 20 states of Mexico: a cross-sectional study," *Scientific Reports*, vol. 9, no. 10094, 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-46543-8

[4] A. Pedroza-Gonzalez, J. Reyes-Reali, M. Campos-Solorzano, E. M. Blancas-Diaz, J. A. Tomas-Morales, A. A. Hernandez-Aparicio, D. M. de Oca-Samperio, E. Garrido, G. S. Garcia-Romo, C. F. Mendez-Catala, P. A. Ortiz, J. S. Ramos, M. I. Mendoza-Ramos, A. D. Saucedo-Campos, and G. Pozo-Molina, "Human papillomavirus infection and seroprevalence among female university students in mexico," *Human Vaccines & Immunotherapeutics*, vol. 18, no. 1, p. 2028514, 2022. [Online]. Available: https://doi.org/10.1080/21645515.2022.2028514

[5] F. J. Solis and L. M. Gonzalez, "A non linear transport-diffusion model for the interactions between immune system cells and hpv-infected cells," *Springer Nature*, vol. 111, no. 16, p. 15557–15571, 2023. [Online]. Available: https://doi.org/10.1007/s11071-023-08616-2