**Materiales y Métodos**

**Descripción del conjunto de datos**

El conjunto de datos utilizado en este estudio proviene de la Dirección General de Epidemiología del Gobierno de México, el cual está disponible como parte de los datos abiertos conforme al Decreto publicado en el Diario Oficial de la Federación el 20 de febrero de 2015. Esta base de datos contiene información epidemiológica relacionada con los casos confirmados de COVID-19 y otros virus respiratorios.

Inicialmente, el conjunto de datos contenía 215,522 registros correspondientes a pacientes con sospecha o confirmación de COVID-19, con un total de 42 columnas que incluyen variables de diversa índole, tales como:

* Datos demográficos (edad, sexo, nacionalidad, etc.).
* Datos clínicos y epidemiológicos (tipo de paciente, fechas clave como fecha de ingreso, fecha de inicio de síntomas y fecha de defunción).
* Variables relacionadas con comorbilidades (diabetes, hipertensión, EPOC, etc.).
* Variables de seguimiento clínico como intubación, estancia en la unidad de cuidados intensivos (UCI) y resultados de pruebas diagnósticas.

El periodo de estudio comprendido abarca desde el inicio de los síntomas (1 de enero de 2024) hasta la fecha de corte más reciente disponible en el conjunto de datos, que corresponde al 3 de marzo de 2025.

**Preprocesamiento de datos**

Inicialmente se seleccionaron los registros confirmados como casos positivos de COVID-19, ya que se busca ver la influencia de las comorbilidades en la defunción de las personas contagiadas por COVID-19 desde el año anterior hasta la actualidad. Posteriormente, se creó una nueva variable para identificar el estado de defunción de cada paciente. Esta variable fue construida con base en la información de la columna que registra la fecha de defunción: si la fecha estaba presente, el paciente se consideró fallecido (valor 1); si la fecha era nula, se consideró que el paciente permanecía con vida al momento del cierre del estudio (valor 0).

Para implementar el modelo de regresión de Cox, se calculó una nueva variable que representa el número de días transcurridos desde la fecha de inicio de síntomas hasta la fecha de defunción (en el caso de los pacientes fallecidos) o hasta el 3 de marzo de 2025 (en el caso de los pacientes que permanecían con vida).

Posteriormente, se seleccionaron las columnas de interés para el análisis, que incluyeron:

* Estado de defunción (variable binaria: 1 = fallecido, 0 = vivo).
* Comorbilidades relevantes para el análisis: neumonía, diabetes, EPOC, asma, inmunosupresión, hipertensión, enfermedades cardiovasculares, obesidad, enfermedad renal crónica y tabaquismo. Todas estas variables fueron codificadas como binarias (1 = presente, 0 = ausente).
* Número de días desde la fecha de inicio de síntomas hasta la fecha del evento (defunción) o el final del periodo de observación.

Durante el proceso de limpieza de datos, se identificaron registros con valores nulos en las variables seleccionadas para el análisis, que presentaban valores 97, 98 y 99 en las variables de comorbilidades, los cuales correspondían a "NO APLICA", "SE IGNORA" o "NO ESPECIFICADO", respectivamente, según el catálogo de variables proporcionado por la Dirección General de Epidemiología. Dado que estos valores representaban una proporción muy pequeña del total de registros, se optó por eliminarlos para evitar la introducción de posibles errores mediante técnicas de imputación.

Tras seleccionar solo los casos confirmados de COVID-19, el conjunto de datos se redujo a 14,684 registros (13,774 vivos y 874 fallecidos). Posteriormente, tras la eliminación de estos registros con datos faltantes o valores inconsistentes, el conjunto de datos final utilizado en el análisis contó con un total de 14,617 registros (13,750 vivos y 867 fallecidos), lo que representó la eliminación del 0.46% de los registros originales (0.17% con respecto a los vivos y 0.8% con respecto a los fallecidos).

### Modelo de Regresión de Cox

El modelo de regresión de Cox, también conocido como modelo de riesgos proporcionales, es una herramienta estadística ampliamente utilizada para analizar datos de supervivencia. Este modelo permite estimar el efecto de múltiples variables explicativas sobre el tiempo hasta la ocurrencia de un evento (en este caso, la defunción).

La función de riesgo del modelo de regresión de Cox se define como:

donde:

* es la tasa de riesgo en el tiempo dada una combinación de variables explicativas.
* es la función de riesgo basal, que representa el riesgo cuando todas las variables explicativas son cero.
* son los coeficientes que estiman el efecto de cada variable explicativa sobre el riesgo.

El modelo de regresión de Cox se basa en el supuesto de **riesgos proporcionales**, que implica que las razones de riesgo entre individuos se mantienen constantes a lo largo del tiempo. Este supuesto se verificó mediante el análisis de residuos de Schoenfeld. En caso de que alguna variable incumpliera este supuesto, se consideró la posibilidad de incluir términos de interacción con el tiempo si fuera necesario, con el fin de modelar el efecto variable de dichas covariables, garantizando que el modelo cumpla con el supuesto fundamental. Se utilizó un umbral de significancia de p < 0.05 tanto en los resultados del modelo como en el test de Schoenfeld.

La implementación del modelo se realizó utilizando el lenguaje de programación R (versión 4.4.1), junto con los paquetes survival (versión 3.6-4) y survminer (versión 0.49), que emplea ggplot2 (versión 3.5.1). Los tiempos de seguimiento se censuraron al 3 de marzo de 2025 para quienes no fallecieron, dado que esta fue la última fecha de actualización del conjunto de datos.

**Resultados**

En este estudio, se analizaron 14,541 registros, de los cuales 13,680 correspondieron a pacientes que sobrevivieron y 861 a pacientes que fallecieron durante el periodo de estudio. Las características clínicas de la población analizada se resumen en la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
| **14541 registros en total** | | **Vivos** | **Fallecidos** |
| **Total (%)** | | 13750 (94.07) | 867 (5.93) |
| **Mujeres (%)** | | 8355 (96.03) | 345 (3.97) |
| **Hombres (%)** | | 5395 (91.18) | 522 (8.82) |
| **Promedio de Edad (std)** | | 43 (± 22) | 65 (± 22) |
| **Tiempos de seguimiento** | **Media (std)** | 285 (± 99) | 11 (± 16) |
| **Mediana** | 265 | 8 |
| **Neumonía (%)** | | 2273 (78.14) | 636 (21.86) |
| **Diabetes (%)** | | 1827 (85.02) | 322 (14.98) |
| **EPOC (%)** | | 361 (80.76) | 86 (19.24) |
| **Asma (%)** | | 472 (95.74) | 21 (4.26) |
| **Inmunosupresión (%)** | | 516 (89.12) | 63 (10.88) |
| **Hipertensión (%)** | | 2428 (86.16) | 390 (13.84) |
| **Enfermedad cardiovascular (%)** | | 440 (82.40) | 94 (17.60) |
| **Obesidad (%)** | | 1128 (92.61) | 90 (7.39) |
| **Enfermedad renal crónica (%)** | | 473 (80.85) | 112 (19.15) |
| **Tabaquismo (%)** | | 590 (86.51) | 92 (13.49) |

La tabla muestra que las mujeres presentaron una menor proporción de fallecimientos (3.97%) en comparación con los hombres (8.82%), lo que sugiere una mayor mortalidad en el grupo masculino.

En cuanto a la edad, se encontró que los pacientes fallecidos tenían en promedio una edad significativamente mayor (65 años ± 22) en comparación con los pacientes que sobrevivieron (43 años ± 22).

Respecto a los tiempos de seguimiento, los pacientes que sobrevivieron presentaron un promedio de 285 días (± 99), mientras que el promedio en los pacientes que fallecieron fue considerablemente menor, con 11 días (± 16.28).

Las comorbilidades más frecuentes en los pacientes fallecidos fueron neumonía (21.86%), enfermedad renal crónica (19.15%), enfermedad pulmonar obstructiva crónica (EPOC) (19.24%) y enfermedad cardiovascular (17.60%). En contraste, condiciones como asma y obesidad presentaron una menor proporción en el grupo de pacientes fallecidos.

### Resultados del modelo de regresión de Cox

|  |  |  |  |
| --- | --- | --- | --- |
| Comorbilidad | Coeficiente | HR (IC 95%) | Valor p |
| Neumonía | 2.25271 | 9.51 (8.13 - 11.13) | < 0.001 |
| Diabetes | 0.48829 | 1.63 (1.38 - 1.92) | < 0.001 |
| EPOC | 0.30938 | 1.36 (1.08 - 1.72) | 0.010 |
| Asma | -0.32631 | 0.72 (0.47 - 1.12) | 0.142 |
| Inmunosupresión | 0.16456 | 1.18 (0.91 - 1.53) | 0.219 |
| Hipertensión | 0.47670 | 1.61 (1.36 - 1.90) | < 0.001 |
| Obesidad | -0.20298 | 0.82 (0.65 - 1.03) | 0.082 |
| Cardiovascular | 0.09750 | 1.10 (0.88 - 1.39) | 0.407 |
| Enf. renal crónica | 0.32608 | 1.39 (1.12 - 1.72) | 0.003 |
| Tabaquismo | 0.34835 | 1.42 (1.13 - 1.78) | 0.003 |

El modelo de regresión de Cox mostró que las comorbilidades como neumonía, diabetes, hipertensión, enfermedadrenalcrónica y tabaquismo se asociaron significativamente con un mayor riesgo de mortalidad por COVID-19 (p < 0**.**05).

En particular, la neumonía presentó el mayor riesgo relativo con un HR = 9.51, lo que indica que los pacientes con esta condición tuvieron aproximadamente 9.5 veces más riesgo de fallecer en comparación con aquellos que no la presentaban. Asimismo, condiciones como la diabetes (HR = 1.63) y la hipertensión (HR = 1.61) también se asociaron con un mayor riesgo de mortalidad.

Por el contrario, ni el asma, ni la inmunosupresión, ni la obesidad, ni las enfermedades cardiovasculares mostraron una asociación significativa con la mortalidad en este análisis (p > 0.05).

El modelo presentó una concordancia de 0.834 con un error estándar de 0.007, lo que indica un buen poder discriminativo del modelo para predecir el riesgo de mortalidad. Además, las pruebas de validación del modelo, incluyendo el Likelihood ratio test (p < 2e-16), el Wald test (p < 2e-16) y el Score (log-rank) test (p < 2e-16) confirmaron la significancia global del modelo, respaldando su solidez estadística.

El test de Schoenfeld fue utilizado para evaluar el cumplimiento del supuesto de riesgos proporcionales en el modelo de regresión de Cox, como muestra la tabla:

|  |  |  |
| --- | --- | --- |
| **Comorbilidad** | **Chisq** | **Valor p** |
| Neumonía | 10.553 | 0.0012 |
| Diabetes | 0.312 | 0.5765 |
| EPOC | 0.024 | 0.8767 |
| Asma | 0.002 | 0.9621 |
| Inmunosupresión | 0.556 | 0.4558 |
| Hipertensión | 0.006 | 0.9408 |
| Obesidad | 0.091 | 0.7625 |
| Enf. cardiovascular | 4.075 | 0.0435 |
| Enf. renal crónica | 1.008 | 0.3153 |
| Tabaquismo | 0.058 | 0.8104 |
| Global | 17.906 | 0.0566 |

El valor de p en el test de Schoenfeld evalúa si existe una relación significativa entre los residuos del modelo y el tiempo, lo que indicaría que el supuesto de riesgos proporcionales no se cumple. Un valor de p < 0.05 indica que una variable incumple este supuesto. En este caso, se identificó que las variables neumonía (p = 0.0012) y enfermedad cardiovascular (p = 0.0435) no cumplieron el supuesto, por lo que se implementaron términos de interacción con el tiempo para estas variables como medida correctiva.

El valor global del test (p = 0.0566) se encontró cercano al umbral de significancia, lo que indica que, en términos generales, el modelo se ajusta razonablemente bien.

**Resultados del modelo de regresión de Cox con términos de interacción**

|  |  |  |  |
| --- | --- | --- | --- |
| Comorbilidad | Coeficiente | HR (IC 95%) | Valor p |
| Neumonía (tt) | 0.96762 | 2.63 (2.44 - 2.83) | < 0.001 |
| Diabetes | 0.52152 | 1.68 (1.43 - 1.99) | < 0.001 |
| EPOC | 0.35032 | 1.42 (1.12 - 1.80) | 0.004 |
| Asma | -0.33384 | 0.72 (0.46 - 1.11) | 0.133 |
| Inmunosupresión | 0.18869 | 1.21 (0.93 - 1.57) | 0.159 |
| Hipertensión | 0.51190 | 1.67 (1.41 - 1.97) | < 0.001 |
| Obesidad | -0.22311 | 0.80 (0.64 - 1.01) | 0.056 |
| Enf. cardiovascular (tt) | 0.06191 | 1.06 (0.97 - 1.17) | 0.203 |
| Enf. renal crónica | 0.35141 | 1.42 (1.15 - 1.76) | 0.001 |
| Tabaquismo | 0.37012 | 1.45 (1.15 - 1.82) | 0.001 |

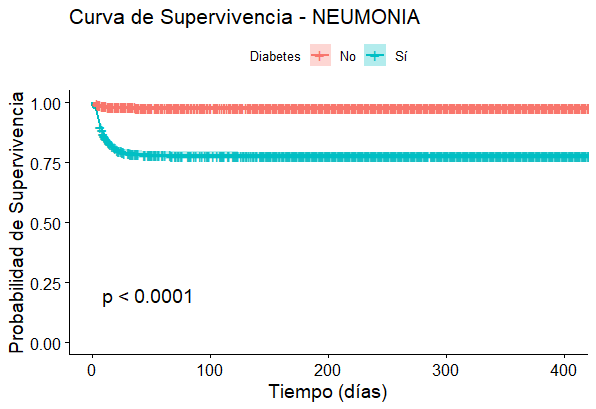
La implementación de los términos de interacción con el tiempo para las variables neumonía y enfermedad cardiovascular permitió corregir el incumplimiento del supuesto de riesgos proporcionales. La neumonía redujo su HR respecto al modelo inicial (pasando de 9.51 a 2.63) aunque continuó siendo significativa (p < 0.001), indicando que el riesgo asociado a esta comorbilidad varía a lo largo del tiempo. Por otro lado, la enfermedad cardiovascular no fue significativa en el modelo inicial ni tras aplicar la corrección (p = 0.203).

Otras variables que continuaron mostrando una asociación significativa con el riesgo de mortalidad fueron diabetes (HR = 1.68), EPOC (HR = 1.42), hipertensión (HR = 1.67), enfermedad renal crónica (HR = 1.42) y tabaquismo (HR = 1.45).

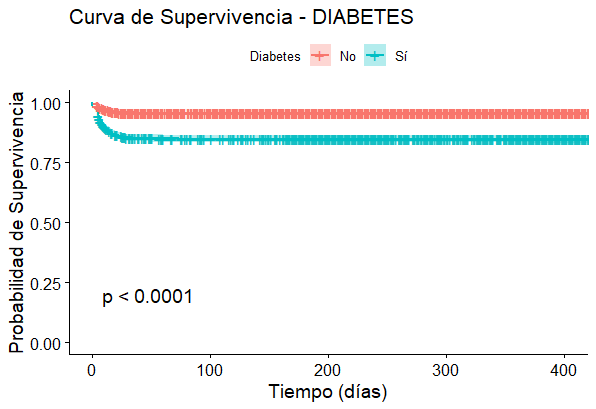
El modelo corregido presentó una concordancia de 0.828 con un error estándar de 0.007, indicando un buen poder predictivo. Las pruebas de validación del modelo, como el Likelihood ratio test (p < 2e-16), el Wald test (p < 2e-16) y el Score (log-rank) test (p < 2e-16) confirmaron la robustez del modelo final.

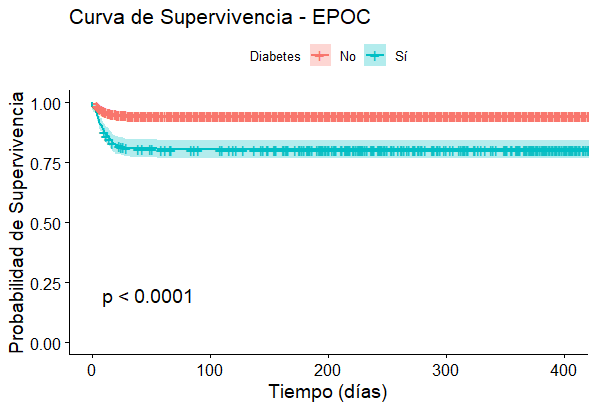
**CURVAS DE KAPLAN-MEIER**

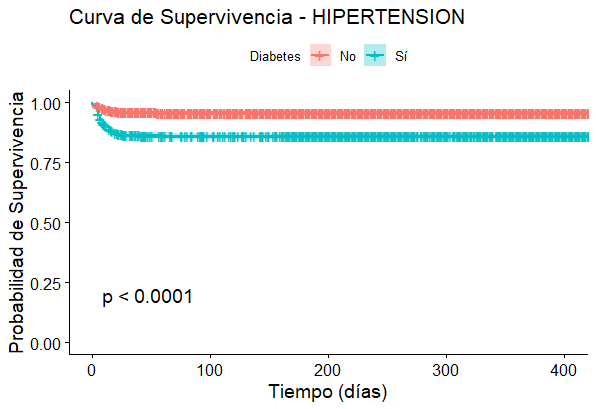
En esta sección se presentan los resultados del análisis de supervivencia mediante curvas de Kaplan-Meier y la prueba log-rank para evaluar las diferencias significativas en la probabilidad de supervivencia entre pacientes con y sin las comorbilidades que resultaron significativas en el modelo de regresión de Cox.



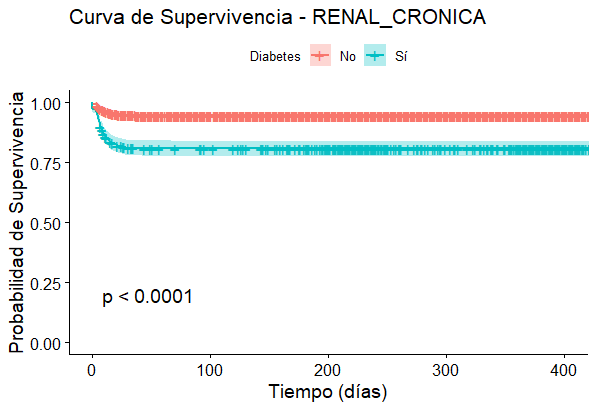
Los pacientes que presentaron neumonía mostraron una probabilidad de supervivencia notablemente menor en comparación con aquellos que no la presentaron. La supervivencia estimada al finalizar el periodo de observación fue del 78.0% en el grupo con neumonía y del 98.0% en el grupo sin esta comorbilidad. La prueba log-rank indicó diferencias altamente significativas (p < 2e-16), confirmando que la presencia de neumonía está asociada a una mayor mortalidad.



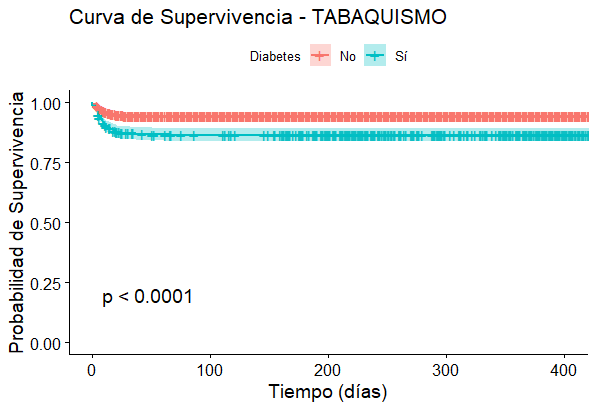
Los pacientes con diabetes mostraron una menor probabilidad de supervivencia en comparación con aquellos que no la presentaron. La supervivencia estimada fue del 84.8% en el grupo con diabetes y del 95.6% en el grupo sin esta comorbilidad. La prueba log-rank resultó significativa (p < 2e-16), evidenciando que la diabetes es un factor de riesgo importante en la reducción de la supervivencia.

La probabilidad de supervivencia en pacientes con EPOC fue del 80.6%, mientras que en aquellos sin esta comorbilidad fue del 94.4% al finalizar el seguimiento. Esta diferencia fue significativa según la prueba log-rank (p < 2e-16), indicando que el EPOC representa un factor de riesgo asociado con una menor supervivencia en pacientes con COVID-19.

Los pacientes con hipertensión presentaron una probabilidad de supervivencia del 86.1%, mientras que en aquellos sin esta condición fue del 95.9%. La prueba log-rank mostró una diferencia significativa (p < 2e-16), destacando que la hipertensión se asocia con una mayor mortalidad en esta población.



Los pacientes con enfermedad renal crónica presentaron una probabilidad de supervivencia del 80.8%, mientras que en aquellos sin esta comorbilidad fue del 94.6%. La prueba log-rank indicó una diferencia significativa (p < 2e-16), confirmando que esta condición está asociada con una menor supervivencia en pacientes con COVID-19.



Los pacientes fumadores presentaron una probabilidad de supervivencia del 86.5%, mientras que en los no fumadores fue del 94.4%. La prueba log-rank mostró una diferencia significativa (p < 2e-16), evidenciando que el tabaquismo es un factor de riesgo asociado con una menor supervivencia en esta población.

Cualquier duda con respecto a la implementación, todo se encuentra en el siguiente link de repositorio: <https://github.com/16danielvm/COX-Regression---Covid-19>

* Conjunto de datos crudo: COVID19Mexico2025.csv
* El preprocesamiento: Preprocesamiento2025.ipynb
* El análisis descriptivo: Preprocesamiento2025\_Analisis descriptivo.ipynb
* Conjunto de datos usado para la implementación: df\_cox2025.csv
* Implementación de modelo, creación de curvas, pruebas de log-rank y test de Schoenfeld: regresion2025.R