ARTÍCULO ORIGINAL

Hacia un modelo predictivo de carácter preventivo del riesgo de infección por COVID-19

*Djamel Toudert**

El Colegio de la Frontera Norte, Departamento de Estudios Urbanos y del Medio Ambiente, Baja California, México

Resumen

Introducción: *La escasez de aplicaciones centradas en la persona y con vistas al desarrollo de la conciencia del riesgo que representa la pandemia de COVID-19 estimula la exploración y creación de herramientas de carácter preventivo accesibles a la población.* **Objetivo:** *Elaboración de un modelo predictivo que permita evaluar el riesgo de letalidad ante infección por el virus SARS-CoV-2.* **Métodos:** *Exploración de datos públicos de 16 000 pacientes positivos a COVID-19, para generar un modelo discriminante eficiente, valorado con una función score y que se expresa mediante un cuestionario autocalificado de interés preventivo.* **Resultados:** *Se obtuvo una función lineal útil con capacidad discriminante de 0.845; la validación interna con bootstrap y la externa, con 25 % de los pacientes de prueba, mostraron diferencias marginales.* **Conclusión:** *El modelo predictivo, basado en 15 preguntas accesibles puede convertirse en una herramienta de prevención estructurada.*

PALABRAS CLAVE: *Modelo predictivo de la letalidad. COVID-19. Herramienta preventiva. México.*

Towards a predictive model for prevention nature of the risk of COVID-19 infection

Abstract

Introduction: *The scarcity of person-centered applications aimed at developing awareness on the risk posed by the COVID-19 pandemic, stimulates the exploration and creation of preventive tools that are accessible to the population.* **Objective:** *To develop a predictive model that allows evaluating the risk of mortality in the event of SARS-CoV-2 virus infection.* **Methods:** *Exploration of public data from 16,000 COVID-19-positive patients to generate an efficient discriminant model, evaluated with a score function and expressed by a self-rated preventive interest questionnaire.* **Results:** *A useful linear function was obtained with a discriminant capacity of 0.845; internal validation with bootstrap and external validation, with 25 % of tested patients showing marginal differences.* **Conclusion:** *The predictive model with statistical support, based on 15 accessible questions, can become a structured prevention tool.*

KEY WORDS: *Predictive model of mortality. COVID-19. Preventive tool. Mexico.*

Correspondencia:
*Djamel Toudert
E-mail: toudert@colef.mx

Fecha de recepción: 02-09-2020
Fecha de aceptación: 02-02-2021
DOI: 10.24875/GMM.20000628

Gac Med Mex. 2021;157:240-245
Disponible en PubMed
www.gacetamedicademexico.com

Introducción

La exacerbación de la crisis provocada por la pandemia de COVID-19 generó una importante demanda de modelos predictivos del riesgo posterior a la infección por SARS-CoV-2.[1] En el marco de este esfuerzo, se concretaron modelaciones predictivas encaminadas a apoyar procesos de planeación y gestión de recursos, afinar protocolos médicos y delinear políticas colectivas de prevención.[1,2] No obstante, la modelación predictiva ha sido escasamente centrada en las personas con miras a generar un mayor involucramiento en la prevención individual consciente. Esta última desempeña un papel trascendente en la estructuración de una prevención colectiva, informada y, sobre todo, colaborativa.[3]

El objetivo de esta investigación fue explorar el universo de datos públicos disponibles sobre COVID-19 en México,[4,5] con la finalidad de generar un modelo predictivo de la letalidad y su implementación en un cuestionario con preguntas accesibles a la población en general. La modelación fue elaborada con una función discriminante de casos positivos a COVID-19, a través del desenlace de curarse o fallecer, otorgando a cada una de las variables predictivas una puntuación por medio de la técnica de *scoring*.[6,7] Para dar mayor difusión a esta herramienta, el cuestionario resultante puede convertirse en una aplicación disponible en línea en la que, además, se desplieguen consejos preventivos personalizados en función de la puntuación que se obtenga.

Métodos

Este estudio utilizó los datos abiertos sobre COVID-19 acumulados al 26 de julio de 2020, publicados y actualizados diariamente por la Dirección General de Epidemiología de la Secretaría de Salud del gobierno federal.[4] También se consideraron los datos abiertos del visualizador analítico para COVID-19 del Instituto Nacional de Estadística y Geografía.[5]

De la base de datos de la Secretaría de Salud[4] fueron seleccionados aleatoriamente 8000 pacientes que permanecieron con vida más de dos meses después de ser positivos a SARS-CoV2 y 8000 que fueron positivos a la misma infección y que fallecieron. De la base de datos del Instituto Nacional de Estadística y Geografía[5] se eligieron los municipios de residencia y sus variables a los cuales correspondían los 16 000 pacientes analizados.

Tomando en cuenta que la mayoría de los datos analizados son expresados en categorías, se realizó la homogeneización a través de una segmentación en cuatro clases incrementales por medio del algoritmo de medias móviles.[8] Los datos resultantes de 75 % de los pacientes positivos sirvieron para calcular el modelo discriminante básico y los del 25 % restante, para llevar a cabo el proceso de validación externa del modelo.

Del conjunto de variables exploradas por su cercanía temática y epistemológica con el estudio, resultaron seleccionadas 15 variables por su carácter discriminante significativo (χ^2 con $p < 0.05$) de los desenlaces de curarse o fallecer[9,10] (Tabla 1).

Enseguida, se realizó un análisis de correspondencias múltiples, conservando los 27 ejes (73.67 % de la varianza total explicada) que resultaron significativos ($p < 0.05$).[6,11] Con estos factores se realizó el cálculo de la función lineal discriminante de Fisher, que permitió generar el modelo de predicción seguido por la atribución de coeficientes a las variables categoriales por medio de una función *score*[7,12] (Tabla 2). La distribución de la puntuación definió la predicción de tres situaciones:

- El pronóstico óptimo de salvarse de la enfermedad.
- La predicción de alerta para pacientes proclives a fallecer.
- La indecisión, que agrupa a individuos positivos con una puntuación intermedia.

La evaluación del modelo desarrollado se realizó por medio de una validación interna con la técnica de *bootstrap* con 1000 remuestreos,[13,14] la cual fue profundizada con el apoyo de la curva ROC (*receiver operating characteristic*) y la gráfica de LIFT.[15] Se llevó a cabo también una validación externa con 25 % de los individuos positivos, población que no participó en la generación del modelo.

Resultados

La evaluación de la calidad y eficiencia del modelo de predicción se llevó a cabo en dos pasos complementarios: la validación interna y la validación externa.

La validación interna del modelo

La validación de la estabilidad interna del modelo evidencia las variaciones mínimas entre el resultado del cálculo básico del modelo y los resultados con *bootstrap*.[16] En todos los casos, esas diferencias

Tabla 1. Características demográficas, clínicas y contextuales de la población analizada*

Características	Vivos (%)	Fallecidos (%)	Total (n = 16 000)
V1. Sexo			
Hombre	26.11	32.41	9363
Mujer	23.89	17.59	6637
V2. Grupos de edad			
1-4	0.30	0.08	60
5-14	0.57	0.04	98
15-24	3.55	0.24	606
25-34	11.61	1.37	2076
35-44	12.38	4.18	2649
45-64	17.13	22.10	6276
≥ 65	4.47	22.00	4235
V3. Embarazo	90.00	10.00	70 (1.05)
V4. Obesidad	43.87	56.13	3476 (21.72)
Comorbilidad**			
V5. Diabetes	24.90	75.10	4061 (25.38)
V6. Pulmonar obstructiva crónica	22.25	77.75	445 (2.78)
V7. Inmunosupresión	29.88	70.12	328 (2.05)
V8. Hipertensión	26.37	73.63	4733 (29.58)
V9. Cardiovascular	23.93	76.07	560 (3.5)
V10. Renal crónica	15.73	84.27	674 (4.21)
V11. Otras	34.62	65.38	621 (3.88)
V12. Pacientes positivos por municipio			
1-948	14.40	16.08	4876
949-3228	15.19	13.61	4608
3229-6832	14.01	14.39	4543
6833-11587	6.40	5.91	1970
V13. Población de 3 a 5 años que asiste a la escuela por municipio			
0-49.36	2.68	3.44	979
49.37-61.76	15.74	16.32	5129
61.77-74.69	23.20	23.16	7417
74.70-100	8.36	7.06	2466
V14. Sector médico en donde se atiende (población 15 894)			
Estatal	1.13	1.15	362
IMSS	15.11	27.46	6765
ISSSTE	1.86	3.72	888
Pemex	0.69	0.75	229
Municipal	0.04	0.04	13
Universitario	0.03	0.03	9
Privada	1.77	0.58	374
Sedena	0.38	0.49	138
Semar	0.52	0.23	118
Secretaría de Salud	28.52	15.50	6997
V15. Nacido en otra entidad que la de residencia	10.99	13.06	3848 (24.05)

*Todas las variables de la tabla son significativas con $p < 0.05$.
**Se consignó en la tabla solo la modalidad afirmativa de las variables comorbilidad, embarazo y obesidad, dado que la negación equivale a cero.
IMSS = Instituto Mexicano del Seguro Social, ISSSTE = Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado, Pemex = Petróleos Mexicanos, Sedena = Secretaría de la Defensa Nacional, Semar = Secretaría de Marina.

fueron inferiores a 0.2 %, lo que indicó una buena estabilidad interna. En el mismo orden de ideas, la curva ROC destacó una puntuación límite de 575, que caracteriza una sensibilidad de 76.6 % y una especificidad de 23.1 %, ambas aceptables en el marco de los objetivos fijados a la presente investigación (Figura 1).

La capacidad discriminante del modelo por medio del área bajo la curva (AUC) fue de 0.845, con un intervalo de confianza de 95 % que indica que la predicción distingue entre curados y fallecidos con una probabilidad de 84.5 %. Un modelo con este valor de AUC es considerado útil para ciertos usos según los criterios de Swets,[17] mientras que es caracterizado como excelente según los criterios de Hosmer y Lemeshow.[18]

Validación externa del modelo

El comparativo de los resultados obtenidos con el modelo básico y el cálculo elaborado con 25 % de los pacientes de prueba dejó entrever diferencias marginales (Tabla 3), las cuales quedaron contenidas en el rango inferior a 0.65 % en valor absoluto, que no parece incidir sensiblemente en la eficiencia del modelo. Lo anterior puede verse en la gráfica LIFT, que revela un modelo capaz de determinar 76.57 % de predicciones acertadas con 50 % de los pacientes positivos (Figura 1).

El resultado de la función discriminante con variables categoriales fue utilizado para elaborar una función *score*; cada uno de los pacientes positivos fue asignado a tres posibles zonas a partir de la puntuación que obtuvieron. En nuestro caso, la puntuación elevada correspondió a la zona de alerta (622-1000) y la puntuación baja coincidió con la zona óptima (0-528). La zona de indecisión identificó a individuos con puntuación oscilante entre 528-622, quienes no pudieron clasificarse en ninguna de dos anteriores (Tabla 4).

Discusión

Por medio de la modelación discriminante se estableció la posibilidad de predicción de curación y fallecimiento al contraer el virus SARS-CoV-2 en México. El modelo de predicción generado tiene la ventaja de estar sustentado en datos públicos accesibles, consta de variables predictivas fáciles de contestar y, por lo tanto, susceptibles de facilitar la

Tabla 2. Validación interna y externa de la función lineal discriminante de Fisher

Elaboración del modelo	Cálculo básico		Cálculo con *bootstrap**	
Grupos discriminados	**Bien clasificados**	**Mal clasificados**	**Bien clasificados**	**Mal clasificados**
Positivos vivos				
n	4492	1508	4481.50	1518.50
%	74.85	25.13	74.69 [0.63]	25.31 [0.63]
Positivos fallecidos				
n	4761	1239	4751.61	1248.39
%	79.35	20.65	79.19 [0.67]	20.81 [0.67]
Total				
n	9253	2747	9233.11	2766.89
%	77.11	22.89	76.94 [0.37]	23.06 [0.37]
Comprobación del modelo	**Cálculo básico**			
Positivos vivos				
n	1502	498		
%	75.10	24.90		
Positivos fallecidos				
n	1574	426		
%	78.70	21.30		
Total				
n	3076	924		
%	76.90	23.10		

*1000 remuestreo aleatorio.
Entre corchete se indica la desviación estándar.

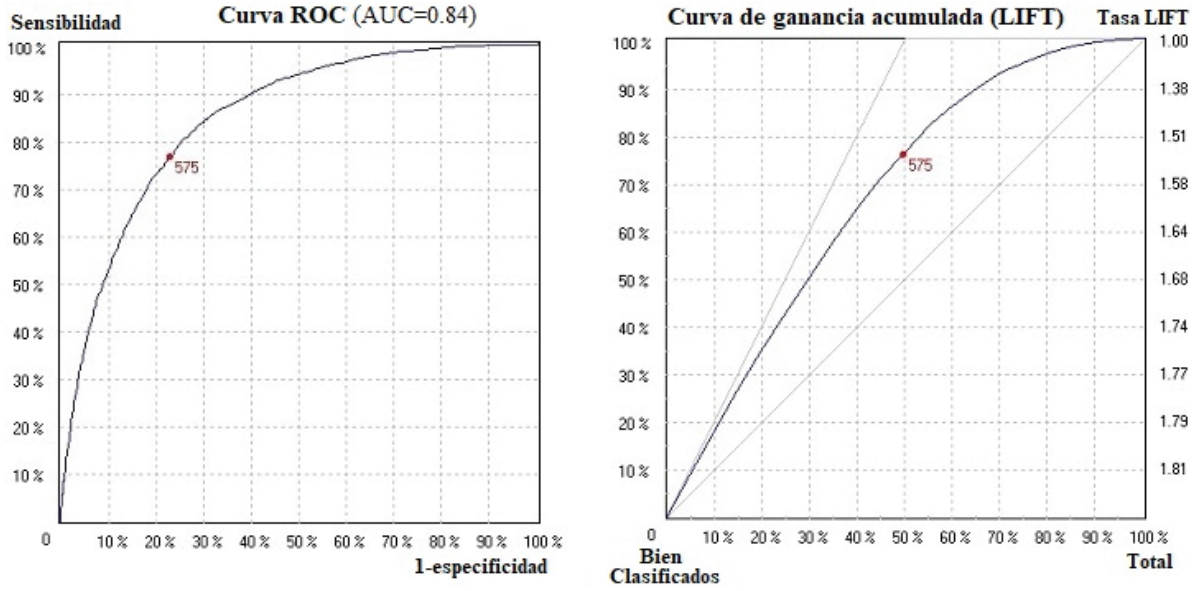

Figura 1. *Calidad y eficiencia del modelo de predicción.*

creación de una herramienta computacional con propósitos preventivos. Durante la crisis de COVID-19, Wynants *et al.*[1] documentaron cerca de 145 modelos predictivos, 34 % dedicados al pronóstico de riesgo de mortalidad o de necesidad de cuidados intensivos. En su mayoría, los modelos que mostraron avances en el conocimiento de la enfermedad[2,19] fueron calificados de sesgados debido a la selección no representativa de pacientes control y al sobreajuste del modelo.[1]

Tabla 3. Coeficientes discriminantes y transformados de las modalidades involucradas

Características	Coeficientes de la función discriminante	Coeficientes transformados
V1. Sexo		
Hombre	1.936	15.85
Mujer	−2.731	0
V2. Grupos de edad (años)		
1-4	-21.688	29.5
5-14	−28.67	5.78
15-24	−30.371	0
25-34	−29.564	2.74
35-44	−18.279	41.08
45-64	4.641	118.95
≥ 65	24.365	185.96
V3. Embarazo	3.323	20.47
V4. Obesidad	1.918	46.73
Comorbilidad		
V5. Diabetes	Municipal	-10.005
V6. Pulmonar obstructiva crónica	−1.38	8.66
V7. Inmunosupresión	3.735	24.7
V8. Hipertensión	3.981	19.37
V9. Cardiovascular	1.918	46.73
V10. Renal crónica	13.754	72.78
V11. Otras	2.487	9.09
V12. Pacientes positivos por municipio		
1-948	3.67	18.6
949-3228	−1.735	0.24
3229-6832	14.01	14.39
6833-11587	−0.901	3.07
V13. Población de 3 a 5 años que asiste a la escuela por municipio		
0-49.36	6.038	27.46
49.37-61.76	0.297	7.96
61.77-74.69	−0.327	5.84
74.70-100	8.36	7.06
V14. Sector médico en donde se atiende		
Estatal	1.718	325.81
IMSS	8.854	350.06
ISSSTE	6.917	343.47
Municipal	−10.005	285.99
Universitario	−18.09	258.52
Pemex	−8.596	290.77
Privada	−21.188	247.99
Sedena	5.287	337.94
Semar	−9.127	288.97
Secretaría de Salud	−8.069	292.56
V15. Nacido en otra entidad que la de residencia		
Sí	1.065	4.76
No	−0.337	0

IMSS = Instituto Mexicano del Seguro Social, ISSSTE = Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado, Pemex = Petróleos Mexicanos, Sedena = Secretaría de la Defensa Nacional, Semar = Secretaría de Marina.

Tabla 4. Distribución de pacientes con una tasa tolerada de error de 10 %

Pacientes positivos	Zona óptima (0-528 puntos)		Zona de indecisión (528-622)		Zona de alerta (622-1000 puntos)	
	n	%	n	%	n	%
Curados	4777	59.71	2423	30.29	800	10
Fallecidos	807	10.09	2992	37.40	4201	52.51
Total	5584	34.90	5415	33.84	5001	31.26

Aunque el presente estudio no reproduce las carencias metodológicas citadas, el modelo propuesto no parece ser el instrumento conveniente para la toma de decisiones de trascendencia operativa en el ámbito médico. Aún con un AUC de 0.845, el modelo propuesto adquiriría más eficiencia con la incorporación de variables sintomáticas y de laboratorio.[2] No obstante, lo anterior alejaría también el modelo de sus objetivos preventivos inducidos por la capacidad de una persona común de contestar instantáneamente un conjunto de preguntas. Tomando en cuenta la escasez de herramientas de pronóstico individual,[1] el dilema en este estudio fue resuelto a favor de los alcances preventivos.

En términos teóricos, el modelo parece corroborar, como se identifica en otras investigaciones, la importancia de la edad, el sexo y la comorbilidad para el riesgo de letalidad.[1,3] Como aportación específica de este estudio, se subraya el peso de factores contextuales como los sectores que atienden a los pacientes con COVID-19, el número de pacientes positivos y la tasa de niños de tres a cinco años que asisten a la escuela por municipio de residencia. La explicación de la incidencia de las dos primeras variables parece evidente, pero en el caso de la última, lo que más se acerca es un referente hipotético de una inmunidad social en esa categoría de infantes.[20]

En términos prácticos, el modelo desarrollado exhibe cualidades estadísticas que le permiten convertirse en una aplicación de carácter preventivo de 15 preguntas estructuradas a partir de las variables de la Tabla 3. En el marco de esta aplicación, se otorgará una puntuación (coeficiente transformado de la Tabla 3) a cada una de las modalidades de respuesta, misma que será sumada, una vez finalizado el cuestionario para la asignación del caso, a una de las tres zonas de decisión (de alerta, de indecisión y

óptima). La asignación de la consulta a una zona de decisión podría alcanzar un nivel de utilidad acompañado de consejos médicos personalizados de carácter preventivo.[21]

Conclusión

Se obtuvo un modelo predictivo con un grado de utilidad que cuenta con un respaldo estadístico y nomológico que permite su conversión en una aplicación de carácter preventivo. Un ejemplo preliminar de esta aplicación puede consultarse en https://bit.ly/2YPlox8

Conflicto de intereses

Los autores declaran no tener conflicto de interés alguno.

Financiamiento

Los autores no recibieron patrocinio para llevar a cabo este artículo.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se realizaron experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Bibliografía

1. Wynants L, van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ. 2020;369:m1328.
2. Ciceria F, Castagnaa A, Rovere-Querinia P, de Cobellia F, Ruggerib A, Gallib L, et al. Early predictors of clinical outcomes of COVID-19 outbreak in Milan, Italy. Clin Immunol. 2020;217:108509.
3. Liu T, Liang W, Zhong H, He J, Chen Z, Guanhao H, et al. Risk factors associated with COVID-19 infection: a retrospective cohort study based on contacts tracing. Emerg Microbes Infec. 2020;9:1546-1553.
4. Dirección General de Epidemiología, Secretaría de Salud; [Internet]. México: Datos abiertos; 2020.
5. Instituto Nacional de Estadística y Geografía [Internet]. México: Visualizador analítico para el COVID-19; 2020.
6. Lebart L, Morineau A, Piron M. Statistique exploratoire multidimensionnelle. Francia: Dunod; 2000.
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-138.
8. Bahmani B, Moseley B, Vattani A, Kumar R, Vassilvitskii S. Scalable k-means++. Proceedings of the VLDB Endowment. 2012;5:622-633.
9. Morineau A. Note sur la caractérisation statistiques d'une classe et les valeurs-tests. Bull Techn Centre Statist Inform Appl. 1984;2:20-27.
10. Nakacha JP, Confais J. Approche pragmatique de la classification. Francia: Technip; 2004.
11. Bardos M. Analyse discriminante. Application au risque et scoring financier. Francia: Dunod; 2001.
12. Droesbeke JJ, Lejeune M, Saporta G, editores. Modèles statistiques explicative pour données qualitatives. Francia: Editions Technip; 2005.
13. Bleeker SE, Moll HA, Steyerberg EW, Donders, G, Derksen-Lubsen DE, Grobbee SE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56:826-832.
14. Wahl S, Boulesteix AL, Zierer A, Thorand B, van de Wiel MA. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Med Res Methodol. 2016;16:144.
15. Vuk M, Curk, T. ROC curve, lift chart and calibration plot. Metodoloski zvezki. 2006;3:89-108.
16. Xie J, Qiu Z. Bootstrap technique for ROC analysis: a stable evaluation of Fisher classifier performance. J Electron. 2007;24:523-527.
17. Swets JA. Measuring the accuracy of diagnosis system. Science. 1988;240:1285-1293.
18. Hosmer DW, Lemeshow S. Applied logistic regression. EE. UU.: John Wiley and Sons; 2000.
19. Bello-Chavolla O, Bahena-López J, Antonio-Villa N, Vargas-Vázquez A, González-Díaz A, Márquez-Salinas C, et al. Predicting mortality due to SARS-CoV-2: a mechanistic score relating obesity and diabetes to COVID-19 outcomes in Mexico. J Clin Endocrinol Metab. 2020;105:1-10.
20. Kloc M, Ghobrial RM, Kuchard E, Lewickie S, Kubiak J. Development of child immunity in the context of COVID-19 pandemic. Clin Immunol. 2020;217:108510.
21. Barajas-Ochoa A, Andrade-Romo J, Ramos-Santillán V. Retos para la educación médica en México en los tiempos del COVID-19. Gac Med Mex. 2020;156:254-257.