Facultad de Ciencias

MODELOS PREDICTIVOS DE MORBI-MORTALIDAD EN PACIENTES CON HEMORRAGIAS INTRACRANEALES

Predictive models of morbi-mortality in patients with intracranial hemorrhages

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCIA DE DATOS

Autor: Pablo Menéndez Fernández-Miranda

Director\es: Lara Lloret Iglesias y Miriam Cobo Cano

Junio, 2023

Agradecimientos

Expreso agradecimiento por este trabajo en especial, a su directora Lara Lloret, por haberme acercado el conocimiento de una disciplina apasionante de la que mi rama de conocimiento se encontraba muy alejada. También quiero agradecerle su dedicación, ayuda y disponibilidad incondicional a cualquier hora y en cualquier día.

Agradezco también a la codirectora, Miriam Cobo, por el apoyo, el soporte, y sus contribuciones a este trabajo.

Por último agradecer a Marta, por todos los días, y a mi familia por ser siempre fuente de apoyo incondicional.

Resumen

Las hemorragias intracraneales (ICH) constituyen una de las principales causas de mortalidad en nuestro medio. Su incidencia es alta, y su letalidad se sitúa en torno al 50%. Las medidas terapéuticas existentes son pocas: soporte vital, tratamiento neuroquirúrgico y tratamiento neurointervencionista. La selección del tratamiento adecuado en cada caso supone una de las principales estrategias para disminuir la morbi-mortalidad, pero también supone un reto. Las escalas y herramientas de estratificación del riesgo son por el momento insatisfactorias, por lo que se precisa de la identificación de nuevos factores pronósticos que sirvan para desarrollar herramientas que ayuden a la toma de decisiones clínicas.

En este sentido, las técnicas estadísticas desarrolladas en el campo del análisis de supervivencia y la reciente adaptación de algoritmos de *Machine Learning* (ML) al manejo de datos censurados, pueden aportar nuevas aproximaciones que permitan el desarrollo de modelos pronósticos eficientes. Partiendo de este contexto, este proyecto se ha propuesto encontrar nuevos factores pronósticos y desarrollar una herramienta de estratificación del riesgo en pacientes con ICH, con la premisa necesaria de que ésta sea de implantación y difusión sencillas.

Para ello, se han recogido datos relativos a más de 160 variables de un total de 300 pacientes. Los resultados apuntaron a los niveles séricos de glucosa ($p < 0.005$), la actividad de la protrombina ($p < 0.005$), las horas de evolución del hematoma ($p < 0.005$), la Escala Coma de Glasgow ($p < 0.005$), y a la comorbilidad asociada, como los factores con mayor capacidad predictiva de supervivencia. En lo referente a los modelos desarrollados, el mayor rendimiento fue obtenido por aquellos basados en la regresión de Cox, que alcanzaron niveles altos de rendimiento con un *c-index* de hasta 0.84 95% IC (0.75,0.90). Los modelos de ML adaptados al análisis de supervivencia, incluyendo *Support Survival Machines*, *Random Survival Forests*, y *modelos Gradient Boosted*, mostraron resultados menos satisfactorios. Con todo ello, se propone un modelo de Cox basado en cinco variables clínicas, como herramienta para ayudar en la toma de decisiones de los pacientes con ICH, a falta de un estudio de validación clínica posterior.

Abstract

Intracranial hemorrhage (ICH) is one of the main death causes in our environment. It has a high incidence and lethality, which is around 50%. The therapeutic measures available are quite limited: life support, neurosurgical treatment and neurointerventional treatment. The selection of the appropriate treatment is one of the main strategies to reduce morbidity and mortality, but it is also a challenge. Scales and tools for the stratification of risk remain unsatisfactory, so new prognostic factors need to be identified in order to develop new models for clinical decision making.

In this sense, the statistical techniques developed in the field of survival analysis and the recent Machine Learning (ML) algorithms adaptation to handle censored data can provide new approaches to develop new efficient prognostic models. In this context, the aim of this project was to find new prognostic factors and develop a risk stratification tool for patients with ICH, with premise of being easy-to-implement and divulge.

To this end, data of more than 160 variables belonging to 300 patients were collected. The results showed that glucose levels ($p < 0.005$), prothrombin activity ($p < 0.005$), hours evolution from the clinical onset ($p < 0.005$), Glasgow Coma Scale ($p < 0.005$), and the comorbidity associated, were considered the factors with the highest predictive power of survival. Regarding the models, the highest performance was obtained by those based on Cox regression, which reached high levels of performance with a *c-index* of up to 0.84 95% CI (0.75,0.90). ML models adapted to survival analysis, including Support Survival Machines, Random Survival Forests, and Gradient Boosted models, showed less satisfactory results. According to the results, a Cox model based on five clinical variables was proposed as a potential tool to aid in decision making for patients with ICH, although further clinical validation study is required.

Índice

Índice de abreviaturas

ICH	Hemorragia intracraneal
GCS	Escala Coma de Glasgow
TC	Tomografía Computerizada
RM	Resonancia Magnética
IA	Inteligencia artificial
ML	*Machine Learning* o aprendizaje automático
CPH	Regresión de Cox o modelo de riesgos proporcionales
AFT	Modelo del tiempo de fallo acelerado
Q-Q	Cuantil-cuantil
EDA	Análisis exploratorio de datos
DE	Desviación típica
IQR	Rango intercuartílico
FDR	*False Discovery Rate*
FWER	*Family wise error rate*
c-index	Índice concordancia de Harrell
CV	*Cross-validation*
IC	Intervalo de confianza
SSVM	*Survival Support Vector Machine*
SVM	*Support Vector Machine*
RSF	*Random Survival Forest*
ST	*Survival Tree*
DT	*Decision Tree*
PCA	Análisis de componentes principales
FA	Análisis factorial

1. INTRODUCCIÓN

1. Introducción

A pesar de los avances médicos y científicos en el campo de la asistencia sanitaria, actualmente persisten algunas enfermedades frente a las que las herramientas terapéuticas son, en un número no desdeñable de casos, incapaces de salvar la vida del paciente. Un ejemplo de este tipo de entidades nosológicas son las hemorragias intracraneales.

1.1. Hemorragias intracraneales

Se define como hemorragia intracraneal (ICH), a cualquier sangrado dentro de la bóveda craneal, bien sea en el parénquima encefálico o en los espacios del sistema ventricular o meníngeos [1]. Es precisamente su localización la que permite realizar la clasificación probablemente más utilizada de las ICH [2] (*Tabla 1*).

Tabla 1. Clasificación de las hemorragias intracraneales.

Tipo	Localización
Hematoma epidural	Espacio epidural
Hematoma subdural	Espacio subdural
Hemorragia subaracnoidea	Espacio subaracnoideo
Hemorragia intraparenquimatosa	Parénquima cerebral, cerebeloso o del tronco

Las hemorragias del sistema ventricular pueden clasificarse de forma conjunta con las hemorragias subaracnoideas, por ser éste una continuidad del espacio subaracnoideo [3].

1.1.1. Epidemiología

Las causas de las ICH son muy diversas, pero a grandes rasgos podemos diferenciar ICH primarias y no primarias. Si bien dentro de las no primarias existen muchas causas, las de origen traumático son las más frecuentes, por lo que en este trabajo hablaremos de ICH primarias y traumáticas. En lo referente a las primeras, las ICH son uno de los dos grandes tipos de ictus, que pueden ser isquémicos o hemorrágicos. Estos últimos constituyen las ICH primarias.

Según datos recientes, la incidencia de ictus en España se sitúa en 291 casos por cada 100.000 habitantes y año [4], de los que aproximadamente un 15% son hemorrágicos. Los ictus representan la segunda causa de muerte, siendo los hemorrágicos mucho más devastadores, con una mortalidad de aproximadamente un 50-60% [5]. Las ICH primarias se diagnostican con mayor frecuencia en mayores de 55 años y en varones, con cierta predilección en población africana y asiática [6].

En cuanto a las ICH traumáticas, la incidencia se sitúa en torno a las 472 hemorragias por cada millón de habitantes y año [7], con un incremento significativo en población mayor de 65 años, que si bien presentan un menor número de sangrados debidos a accidentes de tráfico, ven incrementada la incidencia de caídas con repercusión grave y de sangrados secundarios a la toma de medicación anticoagulante.

1.1.2. Etiología

Como previamente se ha indicado, las ICH pueden ser secundarias a etiologías múltiples, por lo que resulta útil para su descripción la clasificación de las mismas según la localización (*Tabla 1*).

1.1.2.1. Hematoma epidural

Los hematomas epidurales pueden ser tanto arteriales como venosos. Los arteriales representan la amplia mayoría de los mismos, y suelen ser secundarios a traumatismos cerrados sobre la región temporal, que cursan con fracturas que dañan la arteria meníngea media. El resultado del daño arterial es la formación de un hematoma en el espacio epidural [8].

En cuanto a los hematomas epidurales venosos, también suelen ser traumáticos. El mecanismo es el mismo: la rotura de un vaso venoso que ocasiona la repleción del espacio epidural por contenido hemático. A diferencia de los arteriales, los venosos suelen ser más frecuentes en población pediátrica [8].

1.1.2.2. Hematoma subdural

Los hematomas subdurales, a diferencia de los epidurales, suelen ser de origen venoso, sin embargo, al igual que los anteriores, la causa más frecuente es la traumática, pudiendo aparecer también en pacientes bajo tratamiento anticoagulante. En este caso, el espacio que se repleta de sangre es el espacio subdural [9].

1.1.2.3. Hemorragia subaracnoidea

Una hemorragia subaracnoidea es una hemorragia en el espacio subaracnoideo, cuyo origen suele ser de nuevo arterial. Sin embargo, en este caso, además de la etiología traumática, son muy frecuentes las etiologías no traumáticas, destacando la rotura aneurismática. Otras causas no traumáticas y no aneurismáticas incluyen fístulas arterio-venosas y tumores [10].

1.1.2.4. Hemorragia intraparenquimatosa

La hemorragia intraparenquimatosa es una hemorragia en el parénquima cerebral, cerebeloso, o del tronco del encéfalo. Existe una amplia variedad de causas, la mayor parte no traumáticas, entre las que se incluyen la hipertensión, las malformaciones arteriovenosas, la angiopatía amiloide, la rotura aneurismática, tumoral, coagulopática, infecciosa, vasculítica y traumática [8].

1.1.3. Manifestaciones clínicas

Las ICH producen un cuadro clínico neurológico central, pues afectan primariamente al encéfalo. Habitualmente, los síntomas aparecen de forma repentina y evolucionan rápidamente a una mayor severidad. Los síntomas y signos que producen comprenden tanto clínica central común por aumento de la presión intracraneal, como clínica específica por la afectación de estructuras concretas, que dependerá de la localización [11].

En el primer grupo de síntomas se encuentran la cefalea, las náuseas, los vómitos, la inestabilidad, las crisis, y el deterioro del nivel de conciencia, que abarca desde la vigilia hasta el coma. Entre los síntomas específicos asociados a la afectación de localizaciones concretas, destacan la afectación de pares craneales, los déficits motores y sensitivos, las alteraciones pupilares, la hemianopsia, y la desviación de la mirada [11].

1.1.4. Tratamiento

El tratamiento de la ICH comprende tanto terapias médicas, como cirugía abierta, y depende enormemente de la localización del hematoma. En este sentido, es especialmente relevante destacar que la selección adecuada de los pacientes que van a ser sometidos a neurocirugía entraña una de las claves del éxito terapéutico, pues son cirugías de alto riesgo. Operar a pacientes que no deberían de haber sido intervenidos puede conllevar graves secuelas en el paciente, e incluso la muerte, y de la misma forma, no intervenir al paciente que debería ser intervenido puede tener repercusiones similares, graves secuelas o la muerte. Esta selección de candidatos no siempre es sencilla, y en un gran número de casos, aún es controvertida [11].

1.1.4.1. Hematoma epidural

Los hematomas epidurales presentan un pronóstico especialmente devastador, por lo que su tratamiento incluye soporte vital avanzado con control

y monitorización estrecha de la vía respiratoria y de la circulación, y en un gran número de casos, craneotomías descompresivas urgentes para controlar el aumento de la presión intracraneal [8].

1.1.4.2. Hematoma subdural

El tratamiento definitivo de los hematomas subdurales es la evacuación, sin embargo, dependiendo del tamaño y de la localización, pueden ser controlados esperando a su resolución espontánea [8].

1.1.4.3. Hemorragia subaracnoidea

El tratamiento de las hemorragias subaracnoideas depende de su etiología. En el caso de las primarias aneurismáticas, y de las primarias no aneurismáticas debidas a otras lesiones, como malformaciones arterio-venosas, se suele tratar de estabilizar al paciente en el momento agudo para posteriormente abordar, en un segundo tiempo, la causa primaria de la ICH, bien con tratamiento neurointervencionista o bien con tratamiento neuroquirúrgico. En el caso de las hemorragias subaracnoideas traumáticas, los cuidados se centran en el control de las constantes vitales y en la reabsorción espontánea del sangrado [8].

1.1.4.4. Hemorragia intraparenquimatosa

Las hemorragias intraparenquimatosas presentan un tratamiento similar al de las hemorragias subaracnoideas primarias no aneurismáticas. El tratamiento se centra en el control de las constantes, valorando, en función de la clínica y de la localización, la opción de evacuación quirúrgica mediante una craneotomía descompresiva, o un catéter de derivación, así como otras medidas encaminadas a la resolución de las posibles complicaciones.

1.1.5. Pronóstico

La ICH es una de las patologías frecuentes agudas más graves y mortales de nuestro medio. La tasa de mortalidad se sitúa en un 40% al mes y en un 54 % al año. Se calcula que sólo entre el 12% y el 39% de los pacientes logran la independencia funcional a largo plazo [12]. Además, la verdadera eficacia de los nuevos tratamientos con respecto a los cuidados clásicos es controvertida. Algunos estudios relevantes han llegado a señalar que el pronóstico real de la ICH no ha cambiado en los últimos 30 años [13], si bien otros apuntan a que se ha producido un descenso significativo de la mortalidad en las últimas décadas [14].

El pronóstico final es dependiente del nivel socio-económico, así países con altas rentas *per capita* presentan tasas de mortalidad temprana (entre 21 días y 1 mes) de un 25-30%, mientras que la tasa en países de ingresos bajos y medios alcanza el 30-48% [15]. También cabe destacar otro factor altamente relevante: la toma de medicación anticoagulante, especialmente de acenocumarol, que eleva las tasas de mortalidad hasta el 76% [16].

1.1.5.1. Factores pronósticos

En cuanto a los factores pronósticos identificados, podemos distinguir cuatro tipos: detectables en la anamnesis, que incluye la toma de acenocumarol; detectables en la exploración física, que son la puntuación en la Escala de Coma de Glasgow (GCS), y el bajo peso; detectables en el análisis de sangre, que son la presencia de hiperglucemia y la enfermedad renal crónica; y detectables en una prueba de imagen, que hacen referencia a las características del hematoma [12] (*Tabla 2*).

Como puede observarse en la *Tabla 2*, el 50% de los factores pronósticos que han sido establecidos dependen de la realización de una prueba de imagen, bien sea de una Tomografía Computerizada (TC), o bien sea una Resonancia Magnética (RM) [12]. Teniendo en cuenta que la identificación y el tratamiento de los factores pronósticos es fundamental a la hora de estratificar el riesgo de la ICH y poder así tomar la decisión terapéutica más adecuada, principalmente si realizar una intervención quirúrgica o no, parece que resulta necesario dedicar más esfuerzos a encontrar factores pronósticos que no dependan de la realización de pruebas que pueden retrasar la intervención, o no estar disponibles en medios menos desarrollados.

Tabla 2. Factores pronósticos identificados en pacientes con hemorragia intracraneal.

Factor	Tipo
Toma de acenocumarol	Detectable en la anamnesis
Puntuaje bajo en GCS	Detectable en la exploración física
Bajo peso	Detectable en la exploración física
Hiperglucemia	Detectable en el análisis de sangre
Filtrado glomerular <60 ml/min/m^2	Detectable en el análisis de sangre
Volumen de hematoma ≥30 cm^3	Detectable en una prueba de imagen (TC o RM)
Extensión intraventricular	Detectable en una prueba de imagen (TC o RM)
Localización infratentorial	Detectable en una prueba de imagen (TC o RM)
Extravasación de contraste	Detectable en una prueba de imagen (TC o RM)
Lesiones en sustancia blanca	Detectable en una prueba de imagen (TC o RM)

GCS: Escala de Coma de Glasgow; TC: tomografía computerizada; RM: resonancia magnética.

1.2. Análisis de supervivencia

El análisis de supervivencia comprende un conjunto de técnicas estadísticas y de inteligencia artificial (IA), tanto de *Machine Learning* (ML) o aprendizaje automático, como de *Deep Learning* o aprendizaje profundo, que se utilizan para analizar series temporales hasta un evento, conocidas como *time-to-event*. En estos estudios se analiza el intervalo de tiempo (*time-to-event*) que transcurre desde el comienzo del periodo de seguimiento hasta la consecución de un acontecimiento o evento que puede presentarse durante el período de observación, así como el efecto sobre la ocurrencia del mismo de una serie de factores pronósticos o covariables [17,18]. Estas técnicas se desarrollaron originariamente para analizar supervivencia, sin embargo en la actualidad se emplean en múltiples campos, como el financiero o el marketing [17].

1.2.1. Funciones de supervivencia

Los objetivos primordiales de los análisis de supervivencia pasan por la estimación de las funciones de supervivencia y de *hazard* o riesgo, así como por la comparación de estas funciones en diferentes poblaciones de pacientes con el fin de identificar factores pronósticos [19,20].

La función de supervivencia denotada como *S(t)* es una función de probabilidad acumulada que da la probabilidad de supervivencia al evento más allá de un tiempo *t*. Se define a partir de *f(t)*, que es la función que aporta la probabilidad de ocurrencia del evento en un tiempo *t* como [21]:

$$S(t) = P\,(T > t) = \int_t^{\infty} f(u)\,du$$

La función complementaria de *S(t)* es *F(t),* y si bien la primera modeliza la probabilidad de no ocurrencia del evento hasta un tiempo *t*, y por tanto la probabilidad de supervivencia, esta modeliza la probabilidad de ocurrencia del evento hasta un tiempo *t* [21]:

$$F(t) = 1 - S(t)$$

A partir de *F(t)* se puede definir *f(t)*, como [21]:

$$f(t) = F'(t) = \frac{d}{dt} F(t)$$

Del mismo modo, es posible definir *s(t)* a partir de *S(t)*, ya que se trata de su función de densidad [21]:

$$s(t) = S'(t) = \frac{d}{dt} S(t)$$

Por otro lado, se encuentran las funciones de *hazard* o riesgo. *Hazard* se define como el riesgo que tiene un sujeto de padecer el evento en un instante concreto. *Hazard ratio*, que es otro concepto importante a la hora de llevar a cabo un análisis de supervivencia, es un ratio o cociente de *hazards*, y mide cuantas veces más riesgo tiene un sujeto de padecer el evento en presencia o ausencia de un determinado factor. La función que describe el *hazard* o riesgo existente en un tiempo *t* se denota como *h(t)* o *λ (t)* [21]:

$$h(t) = \frac{f(t)}{S(t)}$$

El riesgo acumulado hasta un tiempo *t* o función acumulada de *h(t)* se denota como *H(t)* o *Λ(t)* y matemáticamente se define como [21]:

$$H(t) = \int_0^t h(u)u = \frac{f(t)}{S(t)} = -log\,(S(t))$$

Por último, cabe destacar la relación de todas estas funciones [21]:

$$S(t) = exp[-H(t)] = \frac{f(t)}{h(t)} = 1 - F(t), \qquad t > 0$$

1.2.3. Análisis estadístico

El tratamiento estadístico de las series *time-to-event* comprende algunas particularidades derivadas de la naturaleza de los datos. En primer lugar, la variable objetivo de estudio, dependiente, predictando, o en definitiva, el *outcome*, no es única si no que es doble. Por un lado se define una variable que suele denominarse *Status*, que recoge si se ha producido o no el evento durante el periodo de observación o seguimiento, y por otro lado se define una segunda variable a veces denominada como *Survival (Survival_in_days* en este proyecto*)*, que recoge el tiempo de supervivencia de todos los sujetos [19].

Este diseño estadístico presenta una apreciación importante en cuanto a la variable *Survival*, ya que existen cuatro escenarios posibles para la misma, que se recogen en la *Tabla 3*, y que se pueden clasificar a su vez en dos grupos:

- *Sujetos no censurados*: aquellos que han presentado el evento durante el periodo de seguimiento.
- *Sujetos censurados*: aquellos que no han presentado el evento durante el periodo de seguimiento, bien porque lo presentaron antes (*censura izquierda*), bien porque se han retirado o perdido y no han llegado a completar el periodo de seguimiento (*censura derecha*), o bien porque han completado el periodo de seguimiento libres del evento (*censura derecha).*

Tabla 3. Escenarios posibles de un sujeto en un análisis de supervivencia

Escenario	Evento (*Status*)	Tiempo del evento (*t*)	Seguimiento
No censurado	Sí (1)	Entre t=0 y t=T*	-
Censurado a la derecha	No (0)	-	Completo
Censurado a la derecha	No (0)	-	Parcial
Censurado a la izquierda	Sí (1)	t < t=0	-

*T = tiempo al final del periodo de seguimiento.

Las técnicas estadísticas empleadas para el análisis de supervivencia se caracterizan por ser capaces de lidiar con los datos censurados, es decir, que no precisan de conocer con exactitud el valor de la variable *Survival* para todos los individuos del estudio, permitiendo estudiar tanto observaciones completas o *no censuradas*, como observaciones incompletas o *censuradas*, algo que no permiten otras técnicas como las de comparación de medias.

Cabe destacar que la calidad de las conclusiones de un estudio de supervivencia a menudo puede depender de la cantidad de datos censurados correspondientes a individuos perdidos. Cuanta más proporción de sujetos sin evento pero con seguimiento completo conformen el grupo de datos *censurados*, mayor calidad tendrá el estudio. Las técnicas estadísticas utilizadas para el análisis de supervivencia se describen en el epígrafe *Material y métodos*, y comprenden técnicas de estimación de las curvas de supervivencia y *hazard* no paramétricas, como el *estimador de Kaplan-Meier*, semi-paramétricas, como la *regresión de Cox* o *modelo de riesgos proporcionales* (CPH), y paramétricas, como los *modelos del tiempo de fallo acelerado* (AFT), o el test para la comparación de curvas de varias poblaciones diferentes (*test de Mantel-Cox* o *logrank*).

1.2.4. Inteligencia artificial

El despunte de las técnicas de IA en las últimas décadas ha revolucionado el análisis de datos en prácticamente todos los campos del conocimiento. El análisis de supervivencia no ha sido una excepción, habiendo aparecido progresivamente adaptaciones de varios algoritmos de ML que han tratado de desplazar a la CPH como el *state-of-the-art* en el campo [22].

El *Deep Learning* no se ha quedado atrás. Ya han sido varios los autores que han llevado el prometedor campo de las redes neuronales a la estimación de la supervivencia, con algunos ejemplos como *DeepSurv* [23], *Deep Survival Machines* [24], o *Deep Cox Mixtures* [25], que han logrado mostrar altos niveles de rendimiento, especialmente en presencia de datos complejos como imágenes, o series temporales clínicas con un alto número de covariables.

Una de las principales ventajas que pueden aportar estos modelos basados en IA, es la combinación de aproximaciones. Si bien, como se ha señalado previamente y se verá más adelante en el apartado de *Material y métodos*, las técnicas estadísticas para la estimación de las funciones de supervivencia se basan en un abordaje o bien paramétrico, o bien semi-paramétrico, o bien no paramétrico, los modelos de IA pueden combinar e integrar varias de estas aproximaciones en el mismo modelo, explotando las ventajas de cada una de ellas.

1.3. Contextualización del problema de investigación

Las ICH suponen una de las principales causas de morbilidad y mortalidad en nuestro medio, y a pesar de su gran incidencia e impacto en la salud poblacional, su tratamiento principal, en muchos casos, se basa aún en la aplicación de medidas de soporte vital [26]. La alternativa son los tratamientos neuroquirúrgicos, que se centran en el alivio de la presión intracraneal y en el drenaje del hematoma, pero éstos pueden entrañar elevados riesgos y propiciar graves secuelas en los pacientes llegando, en algunos casos, incluso a producir la muerte.

Por esta razón, una de las principales preocupaciones a la hora de abordar la problemática y el manejo de las ICH pasa por la identificación de perfiles de bajo y alto riesgo. Esto es identificar factores pronósticos o predictivos de morbi-mortalidad, que permitan estratificar a los pacientes en escalas o *scores* de riesgo que puedan servir para seleccionar candidatos y aplicar la intervención más adecuada en cada caso [27,28]. En la línea de esta idea, se pueden encontrar en la literatura varios intentos

de construir modelos predictivos que posibiliten intervenciones prematuras y más precisas [26].

Algunos de estos modelos han demostrado altos niveles de rendimiento en validación interna, por encima del 80% e incluso algunos del 90% [29], sin embargo, muchos de ellos presentan apreciaciones metodológicas relevantes que ponen en duda su generalización y aplicabilidad [26] y que podrían explicar que en la actualidad, ninguno de ellos se haya trasladado a la práctica clínica. Además, prácticamente todos ellos incorporan junto con las variables clínicas, bioquímicas y hematológicas, variables que requieren de la realización de una prueba de imagen y de su interpretación posterior por parte de un radiólogo [26].

A pesar de que este hecho no parece plantear *a priori* un problema en los países desarrollados que cuentan con una amplia disponibilidad de TCs, existen varias desventajas del uso de este tipo de variables. Por un lado cabe destacar que muchas de ellas son variables que, si bien son de fácil interpretación, dependen de la valoración subjetiva del radiólogo lector, véase el efecto de masa de la lesión o la presencia de *spot sign*. Por otro lado, las que si son objetivas, como el volumen del hematoma, pueden presentar una variabilidad inter e intra-observador nada desdeñable. La medición del volumen de la lesión utilizando sus ejes principales puede ser en ocasiones imprecisa, ya que los límites de ésta pueden no estar claramente definidos y depender de la interpretación personal del lector. Si en cambio se utiliza un análisis volumétrico con una segmentación completa tridimensional de la lesión, la medida será más precisa, pero a pesar de que tampoco está exenta de variabilidad inter e intra-observador, requiere del empleo de una cantidad de tiempo que en muchas ocasiones resulta inasumible en un entorno clínico real.

Por último, el empleo de pruebas de imagen encarece la valoración y reduce su aplicabilidad a entornos que cuenten con TCs disponibles en un radio geográfico razonable. El desarrollo de modelos basados en parámetros clínicos o resultado de análisis de sangre sencillos que en muchas ocasiones pueden ser llevados a cabo por instrumentos de medida portátiles y muy poco costosos, como la medición de la glucosa en sangre con un glucómetro, podrían contribuir a mejorar la atención en salud de los pacientes con ICH en cualquier entorno.

Con este pretexto, este trabajo tratará de recoger una muestra de pacientes que permita la construcción de un modelo estadístico o de IA capaz de estratificar el riesgo de pacientes con ICH. El objetivo principal no es construir el modelo con mayor rendimiento de la literatura, si no construir un modelo altamente aplicable en prácticamente cualquier entorno clínico real.

2. HIPÓTESIS Y OBJETIVOS

2. Hipótesis y objetivos

2.1. Hipótesis

1. Es posible detectar factores pronósticos clínicos, bioquímicos y hematológicos en pacientes con ICH utilizando técnicas estadísticas de contraste de hipótesis.
2. Los factores pronósticos clínicos, bioquímicos y hematológicos identificados pueden ser utilizados para construir modelos estadísticos predictivos de supervivencia que alcancen un buen rendimiento.
3. Los factores pronósticos clínicos, bioquímicos y hematológicos identificados pueden ser utilizados para construir modelos predictivos de supervivencia basados en IA que alcancen un buen rendimiento.

2.2. Objetivos

1. Recoger datos de pacientes con ICH para construir una base de datos de tamaño suficiente para extraer conclusiones estadísticas de supervivencia.
2. Anonimizar los datos con técnicas que impidan la desanonimización.
3. Realizar una curación y limpieza de los datos recogidos, haciéndolos fácilmente explotables con técnicas estadísticas y de IA.
4. Realizar un análisis estadístico de los datos aplicando técnicas propias del análisis de supervivencia, con el fin de identificar nuevos factores pronósticos clínicos, bioquímicos y hematológicos.
5. Construir modelos predictivos de supervivencia basados en técnicas estadísticas de análisis de supervivencia.
6. Construir modelos predictivos de supervivencia basados en técnicas de IA adecuadas para el tratamiento de datos de supervivencia.
7. Construir un repositorio abierto con los datos anonimizados, limpiados y curados para que puedan ser reutilizados en otras investigaciones, así como del código utilizado, permitiendo la reproducción del proyecto y favoreciendo la investigación abierta y reproducible.

3. MATERIAL Y MÉTODOS

3. Material y métodos

Para la consecución de este proyecto se recogieron datos pertenecientes a pacientes con ICH, que fueron anonimizados, limpiados, curados y posteriormente analizados estadísticamente. Finalmente, se emplearon modelos estadísticos y de IA específicos para análisis de supervivencia con el objetivo de obtener modelos predictivos de morbi-mortalidad.

3.1. Reclutamiento de pacientes

Se reclutaron pacientes procedentes de tres instituciones sanitarias cuyos nombres no serán revelados por motivos de anonimato y privacidad, especialmente teniendo en cuenta que los datos han sido publicados en un repositorio abierto para potenciar su utilización por la comunidad científica. A lo largo del resto del texto, estos tres centros serán referidos como Institución - 1, Institución - 2, e Institución - 3.

Cabe destacar que los tres hospitales eran centros receptores de urgencias, por lo que las ICH formaban parte de su actividad cotidiana, sin embargo, solamente la Institución - 1 contaba con servicio de neurocirugía y de neurointervencionismo, por lo que los pacientes que precisaron tratamiento neuroquirúrgico o neurointervencionista fueron tratados en la Institución - 1.

Los pacientes fueron reclutados siguiendo una técnica de *muestreo no probabilístico*, el muestreo *por cuotas* (*Figura 1*). Este tipo de muestreo es el equivalente no probabilístico del *muestreo aleatorio estratificado*, y permite asegurar un tamaño muestral adecuado de sujetos representativos al comienzo del estudio [30]. La *cuota* se fijó en base a los datos poblacionales de incidencia [4,7], decidiéndose un tamaño muestral apto para el estudio y plausible para el volumen de pacientes manejados por los tres centros. Este número fue de 300.

A continuación, y teniendo en cuenta que el estudio pretendía analizar supervivencia, se decidió el *periodo de seguimiento* mínimo que debía lograrse para cada sujeto. El tiempo decidido fue de entre cuatro y cinco años.

Posteriormente, se definieron los *criterios de inclusión* y *exclusión*, que se recogen en la *Tabla 4*, y se delimitó la fecha desde la que se debía comenzar el reclutamiento para lograr cumplir la *cuota* fijada y satisfacer los requerimientos relativos al *tiempo de seguimiento*.

Con todo ello, se comenzó reclutando al primer paciente con fecha de admisión el 30-11-2009, incluyendo posteriormente en orden temporal todos los

pacientes con fechas de ingreso posteriores que cumplían los *criterios de inclusión* y *exclusión* hasta alcanzar el tamaño muestral decidido (*Figura 1*). El último paciente reclutado presentó como fecha de ingreso el 18-11-2015.

Tabla 4. Criterios de inclusión y exclusión

Criterios de inclusión
Diagnóstico de ICH.
Anamnesis y exploración física del día del ingreso disponibles.
Análisis de sangre del día del ingreso disponible.
Prueba de imagen del día del ingreso que confirma la presencia de ICH.
Prueba de imagen a los pocos días del ingreso para valorar la evolución.
Anamnesis y exploraciones físicas diarias durante el ingreso para valorar evolución.
Fallecimiento o tiempo de seguimiento completo.
Criterios de exclusión
No diagnóstico de ICH.
Anamnesis y exploración física del día del ingreso incompletas o no disponibles.
No se realiza análisis de sangre el día del ingreso.
No se realiza prueba de imagen el día del ingreso o no se encuentra disponible.
No se realiza prueba de imagen a los pocos días del ingreso para valorar evolución.
No se realizan anamnesis y exploraciones físicas diarias durante el ingreso.
No se dispone de fecha de fallecimiento ni de seguimiento completo.

Para la inclusión de un paciente en el estudio se requirió del cumplimiento de todos los criterios de inclusión y de la no satisfacción de ninguno de los criterios de exclusión.

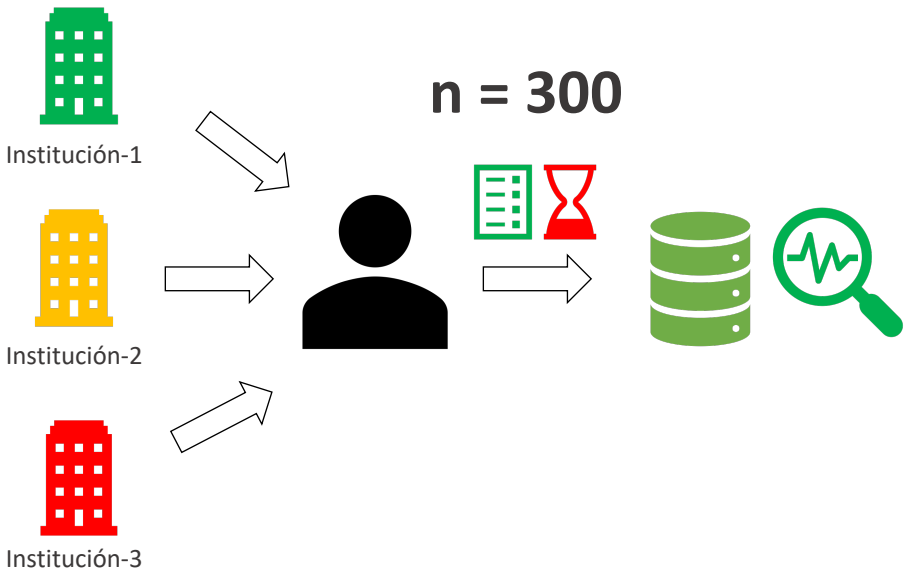


Figura 1. **Muestreo no probabilístico por cuotas.** Se reclutaron de forma retrospectiva pacientes procedentes de tres instituciones sanitarias diferentes, incluyendo, en orden temporal, a todos los pacientes que satisficieron los *criterios de inclusión y exclusión* y contaban con un *periodo de seguimiento* completo hasta alcanzar la *cuota* de 300.

3.2. Aprobación por el Comité de Ética

De acuerdo con lo establecido en las guías de buenas prácticas y calidad científica, este trabajo ha recibido la aprobación del comité de ética correspondiente al ámbito de aplicación. La recogida de datos no comenzó hasta que se contó con la aprobación formal de dicho organismo.

3.3. Protocolo de recogida de datos y datos recopilados

El diseño y la selección de las variables recogidas para este estudio se basó en el conocimiento científico previo, y en la intuición clínica. En total, el número de variables recogidas ascendió a 168, e incluyó dos grandes grupos: *outcomes* y predictores. Los *outcomes* correspondieron con variables relativas al desenlace final del episodio, y delimitaron el objetivo final del proyecto, pues este no fue otro que el de tratar de predecir el valor de los *outcomes* a partir de un grupo de predictores. Los *outcomes* recogidos se resumen en la *Tabla 5*.

Por otro lado, los predictores consistieron en potenciales covariables que se recopilaron con el fin de dotar al *dataset* de información suficiente para predecir los *outcomes*. La mayor parte de los predictores correspondieron con información obtenible en el momento de la llegada del paciente al servicio de urgencias. Dentro de los predictores recogidos se encuentran:

- Identificadores.
- Variables demográficas, como la edad y el sexo.
- Variables relativas al ingreso, entre las que se podrían destacar los días de hospitalización en cuidados intensivos y los días totales de ingreso hospitalario.
- Variables procedentes de la historia clínica, incluyendo los antecedentes personales, la anamnesis, y la exploración física.
- La etiología de la hemorragia, diferenciando entre primaria y traumática y, en el caso de las primarias, identificando la causa específica, a saber, aneurismática, hipertensiva, por angiopatía amiloide, o secundaria a tratamiento fibrinolítico, entre otras.
- Tratamiento neuroquirúrgico o neurointervencionista.
- Análisis de sangre, tanto bioquímico, como hematológico y de coagulación.

En aras de preservar la legibilidad de este documento, la enumeración de las variables que se incluyeron en el estudio se expone en la *Tabla A1* del apartado *Anexo A*. Cabe destacar que estas no son todas las variables que fueron recogidas, si no las

que fueron seleccionadas tras el proceso de limpieza y cura de datos, que ascienden a 143. Para información más detallada y completa acerca de las variables y de sus definiciones, se puede recurrirse a los ficheros de metadatos disponibles en el repositorio del trabajo [31], *ICH_database_anonymized_metadata.csv* e *ICH_database_metadata.csv.*

Tabla 5. Variables *outcome* recogidas

Variable	Definición	Valores
follow_up	Evolución del cuadro	0 (Sin deterioro clínico) 1 (Deterioro clínico debido a la ICH) 2 (Fallecimiento debido a la ICH) 3 (Deterioro clínico no debido a la ICH) NA (Valor perdido)
final_outcome	Desenlace final	0 (Curación completa) 1 (Curación con secuelas) 2 (Fallecimiento hospitalario a causa de la ICH) 3 (Fallecimiento hospitalario por otra causa) 4 (Fallecimiento en los 3 meses tras alta) NA (Valor perdido)
survival_discharge*	Supervivencia al ingreso	0 (No); 1 (Sí)
survival_3d	Supervivencia a los 3 días	0 (No); 1 (Sí)
survival_6d	Supervivencia a los 6 días	0 (No); 1 (Sí)
survival_9d	Supervivencia a los 9 días	0 (No); 1 (Sí)
survival_12d	Supervivencia a los 12 días	0 (No); 1 (Sí)
survival_15d	Supervivencia a los 15 días	0 (No); 1 (Sí)
survival_1m	Supervivencia a los 1 mes	0 (No); 1 (Sí)
survival_3m	Supervivencia a los 3 meses	0 (No); 1 (Sí)
survival_1y	Supervivencia a los 1 año	0 (No); 1 (Sí)
survival_5y	Supervivencia a los 5 años	0 (No); 1 (Sí)
survive	Supervivencia al seguimiento	0 (No); 1 (Sí)
survival_days**	Días de supervivencia	Número de 0 a infinito.
neurosurg	Neurocirugía	0 (No); 1 (Sí); 2 (DVP)
interprocedures	Neurointervencionismo	0 (No); 1 (Sí)

DVP = Válvula de derivación ventrículo-peritoneal. *Este proyecto se centró en el análisis de supervivencia al ingreso, por lo tanto la variable *Status*, que recogió la ocurrencia del evento (muerte), se definió como la inversa de *survival_discharge*. **Por otro lado, *survival_days* dió lugar a la variable *Survival_in_days*.

La recogida de los datos se llevó a cabo utilizando un formulario *Google Forms* con el fin de reducir los errores derivados de la anotación directa en una tabla, así como de reducir el tiempo dedicado a cada caso. Los formularios suponen una herramienta accesible y gratuita que traen consigo algunas ventajas con respecto a la recopilación manual, entre las que se encuentran: la recogida homogénea de los valores de las variables categóricas, evitando que la misma categoría presente al final del periodo de recogida varios valores diferentes en la tabla; la asignación de valores de un sujeto a otro situado en una fila contigua; la asignación de valores de una variable a otra próxima en la tabla; y la reducción del tiempo de recogida.

Los datos fueron recopilados por médicos, lo que resulta un aspecto importante ya que las fuentes de las que se extrajeron (historias clínicas, informes de pruebas de imagen, etc.) contenían textos en jerga médica no estructurados cuya correcta interpretación requería de formación médica especializada. Del total de 300 registros, 250 fueron recopilados por mí, y el resto fueron recogidos por otros dos médicos.

El registro de cada paciente que dejaba el formulario no se encontraba anonimizado, por lo que se seleccionó un terminal en la Institución-1 desde el que se accedió a todas las historias clínicas, incluidas las historias provenientes de la Institución-2 y de la Institución-3, y que almacenó toda la información hasta que ésta fue anonimizada (*Figura 2*).

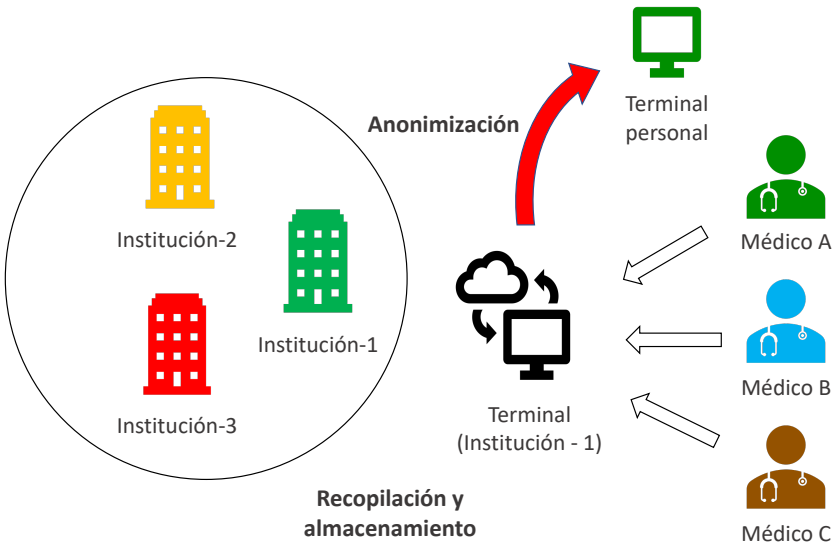


Figura 2. **Protocolo de recogida de datos.** Se recopilaron datos provenientes de tres instituciones, para lo que se seleccionó un terminal localizado en la Institución-1 desde el que se accedió a todas las historias clínicas y donde se almacenaron los datos no anonimizados. Una vez eliminados los identificadores en la primera fase de la anonimización, los datos fueron transferidos a un terminal personal para continuar con el proyecto. Los datos fueron recogidos por tres médicos.

3.4. Preprocesamiento de datos

Una vez finalizada la recopilación de los datos, se procedió a realizar un tratamiento inicial con el objetivo de transformar la base de datos inicial en una base de datos directamente procesable estadísticamente y a través de técnicas de ML. También se prestó especial dedicación a la anonimización completa de la misma, suprimiendo cualquier información que pudiera posibilitar su desanonimización.

3.4.1. Anonimización

Como es bien sabido, los datos médicos son especialmente sensibles y su tratamiento y divulgación se encuentran regulados bajo el Reglamento General de Protección de Datos, cuyo objetivo primordial es garantizar la privacidad de los sujetos [32]. El primer paso para dar cumplimento a la normativa y dotar a los datos de privacidad, es la eliminación de los identificadores. De esta forma, la primera acción que se llevó a cabo en este proyecto una vez culminada la recopilación de los datos, fue la supresión de toda la información que pudiera permitir identificar a un sujeto concreto. Esta acción se llevó a cabo en el terminal ubicado en la Institución-1, que fue utilizado para recoger y almacenar los datos. La consecución de esta primera fase de anonimización permitió el traslado del *dataset* a un terminal personal, desde el que se realizaron el resto de tareas de este proyecto.

En un segundo tiempo se procedió a la anonimización de aquellas variables que si bien por si solas no permitirían identificar sujetos específicos, por agregación podrían llegar a funcionar como identificadores, los llamados cuasi-identificadores. Dentro de las técnicas de anonimización más utilizadas para lidiar con los cuasi-identificadores, se encuentra la generalización. Esta técnica consiste en la transformación de los valores de las variables que forman los cuasi-identificadores en otros valores que presenten menor cardinalidad y precisión que los datos de entrada, volviendo a la base k-anónima con respecto a dicho atributo.

Siguiendo una estrategia similar a la anterior, se procedió al tratamiento de varios de los cuasi-identificadores que presentaba la base de datos, en concreto las fechas de nacimiento, mortalidad, admisión hospitalaria, realización de análisis de sangre, realización de TC y alta. Estas variables combinadas entre sí, con otros datos clínicos, o con otros datos externos podrían llegar a permitir la desanonimización del *dataset*, por lo que se optó por eliminarlas. Sin embargo, siguiendo los principios de la generalización, no se realizó exclusivamente una deleción de la información, si no que se definieron tres nuevas variables que

permitieron preservar toda la información útil para el análisis que contenían los atributos fecha, pero eliminaron la información relativa a la localización del sujeto en un punto concreto del tiempo.

Las tres nuevas variables que surgieron del proceso de anonimización fueron: *survival_days,* que recoge los días de supervivencia de un paciente desde la fecha de ingreso hasta el final del periodo de seguimiento; *age*, que contiene las edades de los pacientes en el momento que acudieron a urgencias; y *time_between_CT_bloodanalysis*, con los tiempos que pasaron desde que el paciente se realizó el TC hasta que se realizó el análisis de sangre.

Como se puede constatar, la nueva base de datos sin las variables fecha y con estas tres nuevas variables, mantiene la información relativa a la cantidad de tiempo y elimina las coordenadas temporales que permitían localizar al sujeto en un punto concreto del tiempo. De esta forma se preserva toda la información relevante para el análisis y se elimina la mayor parte de la información responsable de la pérdida de privacidad.

El resto de los atributos de la base de datos no han sido sometidos a procesos de anonimización, pues se ha considerado que son atributos sin riesgo.

3.4.2. Limpieza y curación

Se entiende por limpieza y curación de datos al preprocesamiento encaminado a organizar y depurar la información contenida en una base de datos de acuerdo con las mejores prácticas, y con el fin de optimizar su explotación, conservación y reutilización. Por lo tanto, se incluyen en esta definición todos los tratamientos dirigidos a la resolución de los problemas intrínsecos que presente el *dataset* en lo que a almacenamiento de la información se refiere.

La limpieza y curación de datos es una fase necesaria en las investigaciones que se desenvuelven en el dominio de la ciencia de datos, y cuya necesidad surge de la imposibilidad de realizar una recopilación perfecta de la información. Dentro de las acciones comúnmente llevadas a cabo en esta etapa, se encuentran la identificación y corrección de datos incoherentes, el tratamiento de los valores perdidos, y la asignación de los tipos de datos adecuados para cada atributo [33].

En esta línea, es importante hacer una mención al manejo de los datos extremos. Una técnica frecuentemente utilizada para detectar datos que han sido erróneamente recogidos es prestar atención al rango de una variable y comprobar que los límites del mismo no exceden los valores que podrían

considerarse esperados. Cabe realizar en este sentido una importante aclaración que es fuente de error en algunos trabajos: la diferenciación entre *outliner* y dato erróneamente recogido. Mientras que los primeros forman parte de la distribución observada de la variable y no deben ser eliminados, los segundos deben de ser corregidos. No obstante, hay que tener en cuenta que los *outliners* pueden llegar a tener gran impacto sobre los estadísticos muestrales (aumentan la desviación típica, la asimetría y la curtosis, entre otros), por lo que su identificación es de gran importancia para adaptar el análisis estadístico posterior.

En lo referente a este proyecto, la tarea de limpieza y curación se realizó después de la primera fase de anonimización: la de eliminación de los identificadores. En primer lugar, se comenzó con la detección y corrección de valores erróneamente recogidos, para lo que se utilizaron tanto análisis de rangos, como pruebas gráficas para la visualización de las distribuciones, destacando gráficos de barras, histogramas, y gráficos cuantil-cuantil (Q-Q). Un ejemplo representativo de este proceso fue la detección de un valor de potasio de 54 mEq/L que, teniendo en cuenta que los valores de potasio fisiológicos oscilan entre 3.5 y 5.5 mEq/L [34] y que una elevación por encima de 7.5 mEq/L puede ser incompatible con la vida, permitió asegurar que el dato había sido erróneamente recogido, probablemente por no haber el separador decimal.

El análisis de rangos y la visualización de las distribuciones permitió también la identificación de algunas variables que presentaban el mismo valor, o bien para todos los sujetos, o bien para la gran mayoría de ellos, imposibilitando la extracción de conclusiones por falta de diversidad suficiente en los valores del atributo. Estas variables se consideraron inútiles para la investigación y fueron eliminadas. Algunos ejemplos de variables eliminadas fueron la toma de *tiazolidinedionas* y la toma de *gliflozinas*, pues ninguno o prácticamente ningún paciente las tomaba. El resto de variables eliminadas en este proceso se encuentra detallado en el repositorio del trabajo [31].

A continuación se realizó una inspección visual de la base de datos detectando atributos que recogían la misma información. Si bien este tipo de análisis es propio de los análisis de correlación, se llevó a cabo en dos fases. En un primer momento se eliminaron las variables cuya información era idéntica por compartir la misma definición o estar una incluida en la otra, y posteriormente se estudió la correlación bivariada entre las variables cuantitativas. Sin embargo, la primera fase se realizó en la etapa de limpieza y curado de datos, y la segunda en la de análisis estadístico inferencial. El objetivo fue el de eliminar información redundante.

En relación a este último punto, podría plantearse la pregunta de cómo pudo ocurrir que hubiera variables que recogían la misma información, habiéndose realizado un buen diseño de los datos de recogida. Un ejemplo representativo de esta situación fue lo ocurrido con la diabetes. Se recogió la presencia de diabetes en dos variables, una dicotómica que indicaba la existencia (1) o ausencia de la enfermedad (0), y otra numérica que registraba el tipo de diabetes que presentaba el paciente. Al final de la recogida de datos, se observó que prácticamente todos los diabéticos eran diabéticos tipo II, por lo que se podía asumir que todo el que tenía 0 como valor de la primera variable no era diabético, mientras que todo el que tenía 1 era diabético, y además, diabético de tipo II.

Posteriormente se continuó el proceso de limpieza y curación con la reasignación de los nombres de las variables. Las premisas para seleccionar buenas denominaciones difieren si se desea diseñar un formulario para la recopilación de datos con respecto a si se desea manejar el *dataset* construido para el análisis. Cuando se decidieron los nombres de las variables para el formulario, el objetivo primordial era preservar la interpretabilidad por parte del médico que cumplimentaría la hoja de recogida. En este sentido, los nombres completos, largos y explicativos, como "Fumador (paquetes/año: cigarrillos día/20 x años que lleva fumando; 1234 si no hay datos)", reducen los errores a la hora de introducir un registro. Sin embargo, el manejo de las variables durante el análisis estadístico y la construcción de modelos de IA utilizando código de programación demanda nombres cortos, claros, y concisos, como "n_tobacco", que fue la denominación que más tarde sustituyó al nombre anterior de la variable que registraba la cantidad de tabaco que fumaba el paciente. La selección apropiada de los nombres fue una etapa a la que se le dedicó esfuerzo y tiempo.

Por último, una vez que los datos se habían limpiado y curado, se almacenaron en formatos que permitieran cargar el *dataset* y preservar todo el trabajo realizado, especialmente que mantuvieran los tipos de datos asignados. Los formatos escogidos fueron HDF5 [35] para Python, RDS [36] para R.

3.5. Análisis estadístico

El análisis estadístico es una parte fundamental de cualquier estudio de investigación, especialmente si se enmarca en el ámbito médico. En este proyecto, el análisis estadístico se llevó a cabo en tres etapas: se realizó un análisis descriptivo de las variables recogidas; a continuación se realizó un análisis inferencial utilizando contrastes de hipótesis; y por último se llevó a cabo un análisis de supervivencia bajo una aproximación estadística.

3.5.1. Estadística descriptiva y EDA

En las investigaciones médicas, la estadística descriptiva conforma la aproximación inicial del análisis de datos. Su objetivo es aportar un resumen de la información analizada a través del cálculo de un conjunto de estadísticos y de la visualización de las distribuciones empleando técnicas gráficas adecuadas. Estos procedimientos deberán ser capaces de reflejar los patrones y tendencias presentes en los datos, en ocasiones imperceptibles a primera vista [37]. En el campo de la ciencia de datos este tipo de análisis se incluye con frecuencia en lo que se denomina comúnmente *análisis de datos exploratorio* (EDA).

Las variables cuantitativas se estudiaron calculando los siguientes estadísticos descriptivos: media, desviación típica (DE), mediana, percentiles 25 y 75 que permitieron calcular el rango intercuartílico (IQR), mínimo y máximo para calcular el rango, asimetría, y curtosis. La definición de curtosis escogida fue la definición de Pearson, que establece como mesocurtosis el valor igual a tres. La visualización de las variables cuantitativas se realizó con diagramas de barras en el caso de variables discretas, y con histogramas en el caso de variables continuas.

Para las variables cualitativas se extrajeron tablas de frecuencia absoluta y relativa. La técnica de visualización empleada para las distribuciones categóricas fue el diagrama de barras. Cabe destacar que las variables *Likert*, como el GCS, recibieron el tratamiento propio de variables categóricas, tal y como se recomienda en la literatura [38]. Los atributos referentes a fechas fueron descritos aportando dos estadísticos (fecha menor y fecha mayor), la tabla de frecuencias absolutas por años, y la representación mediante diagramas de barras.

Por último, se realizaron análisis descriptivos conjuntos de las variables etiquetas como *outcome*, junto con algunas variables predictoras que bajo la intuición clínica podían presentar asociaciones estadísticamente significativas con el objetivo de guiar el análisis inferencial posterior. Para ello se recurrió a la agrupación por subpoblaciones, a la utilización de tablas pivotantes, y a la representación mediante diagramas de barras, gráficos de dispersión, y gráficos de cajas. Cabe destacar que este tipo de análisis conjuntos tenía una finalidad meramente exploratoria, y en ningún caso los resultados fueron utilizados para confirmar la existencia de asociaciones estadísticamente significativas.

En los sucesivos apartados del texto, la presentación de los estadísticos descriptivos de las variables cuantitativas seguirá la forma media ± DE o mediana ± IQR según proceda, y la de las variables cualitativas frecuencia absoluta n (%).

3.5.2. Estadística inferencial

La estadística inferencial agrupa un conjunto de técnicas que tratan de extraer conclusiones de los datos. A diferencia de la estadística descriptiva, que no infiere si no que describe, la estadística inferencial permite realizar inferencias [39].

El análisis inferencial comenzó con el estudio del supuesto de normalidad, que establece que los datos siguen una distribución *gaussiana*. Su valoración es fundamental a la hora de orientar el análisis inferencial posterior, e incluso de interpretar los estadísticos descriptivos. En las distribuciones que siguen una morfología normal, la media y la desviación típica captan adecuadamente la tendencia central de la distribución y su dispersión, lo que permite la aplicación de un grupo de test estadísticos denominados test paramétricos, que presentan alta sensibilidad a la hora de detectar asociaciones estadísticas [40].

Por el contrario, distribuciones que vulneran el supuesto de normalidad deben acogerse a otro conjunto de test menos sensibles, los test no paramétricos [40], basados habitualmente en otros estadísticos que buscan ser capaces de captar la tendencia central y la dispersión de la muestra, a pesar de presentar alteraciones en la asimetría y curtosis. Con frecuencia, estos estadísticos son la mediana y el IQR.

No se debe olvidar que en el caso de comparación de muestras de varios grupos, solo podrán ser aplicados los test paramétricos en el caso de que todas las muestras cumplan el supuesto de normalidad, en caso contrario deberá recurrirse a la estadística no paramétrica (*Tabla 6*).

En este proyecto el supuesto de normalidad global de las variables se estudió utilizando el estadístico de *Shapiro-Wilk*, y la inspección visual con gráficos Q-Q. Además, se comprobó la simetría y mesocurtosis de todas las variables. De acuerdo con la literatura, el estadístico de *Shapiro-Wilk* puede resultar poco preciso para muestras mayores de cincuenta [41], mientras que los gráficos Q-Q constituyen una de las mejores herramientas a la hora de valorar normalidad en distribuciones con tamaño muestral grande, de varios cientos de pacientes [42].

La comprobación de la normalidad de dos o más muestras previamente a su comparación mediante contrastes de hipótesis, se realizó utilizando gráficos de violín. Este gráfico es actualmente reconocido como uno de los mejores para valorar conjuntamente normalidad, morfología y dispersión de una distribución,

que son los tres atributos fundamentales a la hora de decidir el test estadístico más adecuado.

Tras el análisis de normalidad, se llevó a cabo un análisis de correlación de las variables cuantitativas utilizando los coeficientes de correlación *r* de Pearson (variables normales) y *ρ* de Spearman (variables no normales), y se realizaron contrastes de hipótesis múltiples con el objetivo de identificar factores de riesgo. Los contrastes de hipótesis realizados se encuentran detallados en el repositorio del trabajo [31], y los criterios de decisión en base a los cuales se decidió el test estadístico adecuado se resumen en la *Tabla 6*.

Tabla 6. Test estadísticos bivariados para muestras independientes [43].

		Supuesto		
Test	Muestras	Normalidad	Homocedasticidad	Forma
Cuantitativa vs cuantitativa				
Correlación *r* de *Pearson*	2	Sí	Sí	Sí
Correlación *ρ* de *Spearman*	2	No	No	No
Cuantitativa vs cualitativa				
T de *Student*	2	Sí	Sí	Sí
Test de Welch	2	Sí	No	Sí
ANOVA	≥ 3	Sí*	Sí	Sí
U Mann-Whitney-Wilcoxon	2	No	Sí**	Sí**
H de Kruskal-Wallis	≥ 3	No	Sí	Sí
Test de la mediana de *Mood*	≥ 2	No	No	No
Cualitativa vs cualitativa				
Test de χ^2 con SMC	≥ 2	No	No	No

SMC: Simulación de Monte Carlo. *Normalidad de los residuales. **Si se respetan los supuestos el test de *Mann-Whitney* evaluará si existen diferencias entre las medianas de dos distribuciones, no obstante también puede emplearse sin igualdad de varianzas ni de forma, pero en este caso solo informará de si existen diferencias entre dos distribuciones, no diferencia de medias.

Se determinó un nivel de significación *α* de 0.05 para alcanzar la significación estadística en los contrastes de hipótesis, y un coeficiente de correlación *r* o *ρ* de 0.5 en los análisis de correlación para considerar relevante el hallazgo, siempre y cuando este hubiera demostrado ser significativo.

Debido al alto riesgo de cometer *errores α* o *tipo I* secundarios a la comparación múltiple, se consideró necesario aplicar una técnica de corrección de los valores de *p*. La técnica escogida fue la llamada *False Discovery Rate* (FDR), un método diseñado para controlar la tasa de falsos positivos en un análisis de

comparaciones múltiples [44]. La selección de este método en virtud de otras técnicas como las del grupo *family wise error rate* (FWER), entre las que se encuentra el popular método de *Bonferroni*, radica en que FDR es menos conservador y presenta mayor potencia estadística que los métodos FWER, una característica que suele hacer a esta técnica de elección en las investigaciones médicas [45].

Todos los test que resultaron estadísticamente significativos fueron sometidos a la comprobación visual de los resultados utilizando un gráfico adecuado: gráficos de dispersión para la comparación de dos variables cuantitativas; histogramas, gráficos de cajas, y gráficos de violín para la comparación de una variable cuantitativa y otra cualitativa; y diagramas de barras para la comparación de dos variables categóricas.

3.5.3. Análisis de supervivencia: aproximación estadística

Como se indicó previamente en el epígrafe *Introducción*, el análisis de supervivencia hace referencia a un conjunto de técnicas de análisis de datos especializadas en el manejo de series temporales hasta la ocurrencia de un evento (*time-to-event*). La principal ventaja de estos métodos radica en su capacidad para manejar los datos propios de un estudio de supervivencia: observaciones *censuradas* y *no censuradas*.

Si bien en el análisis estadístico inferencial se evaluaron asociaciones múltiples entre los *outcomes* y diferentes factores predictores, el análisis de supervivencia de este trabajo, que representa el núcleo central del mismo, se centró en el estudio de la supervivencia a la ICH. Por lo tanto, el evento estudiado fue la muerte.

La variable que recogió la información relativa a la supervivencia al episodio fue *survival_discharge*, pues describe si el paciente fue alta al domicilio y por tanto se consideró medicamente fuera de riesgo, o si falleció durante el ingreso.

Se definieron las variables principales de supervivencia:

- *Status*: recoge la ocurrencia o no del evento (muerte) y fue definida como la inversa de *survival_discharge*, tomando los valores *True* o 1 si ocurrió el evento, y *False* o 0 si no fue observado.
- *Survival_in_days*: recoge los días de supervivencia, y en el dataset se correspondió con la variable *survival_days*.

A continuación, se procedió a la estimación de las curvas de supervivencia, *hazard*, y *hazard* acumulado, *S(t)*, *h(t),* y *H(t)*, respectivamente, lo que conforma el eje central de los análisis de supervivencia. Estas funciones permiten no solo modelizar la serie temporal y hacer predicciones, sino también estudiar la influencia de covariables sobre la ocurrencia del evento con el fin de descubrir nuevos factores pronóstico.

3.5.3.1. Estimador de Kaplan-Meier y test Mantel-Cox o *logrank*

Una de las técnicas más empleadas para estimar la función de supervivencia *S(t)* es el *estimador de Kaplan-Meier* [46]. Se trata de una aproximación no paramétrica que da como resultado una función escalonada en la que los escalones corresponden con los momentos t_i en los que han ocurrido uno o más eventos. Matemáticamente se define como:

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

- n_i: número de pacientes en riesgo inmediatamente antes del tiempo t_i.
- d_i: número de eventos ocurridos en el tiempo t_i.

Con frecuencia el método de *Kaplan-Meier* se combina con el test de *Mantel-Cox* o *logrank* [47], un test estadístico diseñado para la comparación de curvas de supervivencia y cuyos supuestos de aplicación son compartidos con los del estimador de *Kaplan-Meier*: la censura no está relacionada con el evento; la probabilidad de supervivencia no depende del momento en el que se produjo el reclutamiento; y los eventos ocurrieron en los tiempos especificados.

Teniendo en cuenta que los supuestos de aplicación de éstos métodos se cumplían en nuestro estudio, se llevó a cabo una estimación de las curvas de supervivencia global y de subpoblaciones definidas por la presencia o ausencia de algún potencial factor pronóstico correspondiente con una variable categórica, utilizando el *estimador de Kaplan-Meier*. Posteriormente, se compararon las curvas utilizando el test *logrank* y ajustando los valores de *p* mediante FDR. Cabe destacar que las variables de agrupación utilizadas fueron variables categóricas en todos los casos, ya que la dicotomización de variables cuantitativas para separar subpoblaciones de supervivencia no se considera una buena práctica. La aplicación del *estimador de Kaplan-Meier* junto con el test de *Mantel-Cox* debe reservarse únicamente para el estudio de covariables categóricas.

3.5.3.2. Regresión de Cox o modelo de riesgos proporcionales

Surge entonces la necesidad de contar con un método capaz de analizar el efecto de covariables cuantitativas sobre las curvas de supervivencia y la ocurrencia del evento. Este método es la *regresión de Cox* o *modelo de riesgos* proporcionales, que además de permitir el análisis de cofactores cuantitativos, permite llevar a cabo análisis multivariantes, no quedando restringido a la valoración bivariada que permitía el test *logrank* [48].

Cabe destacar que a diferencia del *estimador Kaplan-Meier* que se configura como una solución no paramétrica, el modelo CPH aporta una solución semi-paramétrica aplicable cuando se cumplen los siguientes supuestos: debe existir independencia de los tiempos de supervivencia entre los individuos de la muestra; la relación entre los predictores y el logaritmo del *hazard* (coeficientes) debe de ser lineal; debe existir proporcionalidad de los riesgos, es decir, el *hazard ratio* debe ser constante a lo largo del tiempo.

La regresión de Cox se describe matemáticamente a través de la siguiente fórmula:

$$h(t|x_i) = h_0(t)exp^{(\beta' x_i)}$$

- $\boldsymbol{h(t|x_i)}$**:** función hazard.
- $\boldsymbol{h_0(t)}$: función hazard basal.
- $\boldsymbol{exp^{(\beta' x_i)}}$: hazard ratios.

Como todo modelo de regresión, un aspecto clave en el ajuste de un regresor de CPH es la selección de las variables. El número aproximado de variables a incluir en el modelo se estimó basándose en la *regla de los 10*, por la que se establece que la relación entre el número de características y el tamaño muestral debe situarse en torno a 10:1, para evitar el *overfitting*.

En cuanto a la selección de los atributos, se comenzó llevando a cabo una regresión de Cox univariante para cada variable y ranqueando posteriormente los *scores* obtenidos por cada modelo. Las métricas de evaluación utilizadas fueron el índice de concordancia de Harrell (*c-index*) y el índice de concordancia de Uno (*Anexo B*). Este método permitió identificar las covariables con mayor poder predictivo. A continuación, se seleccionaron las variables que si bien no habían obtenido los *scores* más altos en el análisis univariante, presentaban alta sospecha clínica de poder influir en la ocurrencia del evento.

Con todo ello se ajustaron cuatro regresores CPH mediante *K-fold Cross-validation* (CV), siguiendo el procedimiento que se detalla en el *Anexo B*. Los dos primeros se ajustaron utilizando exclusivamente las variables con mayor significación en las regresiones univariantes; el tercero se ajustó con las variables que habían sido seleccionadas por su significado clínico; y el cuarto se ajustó utilizando tanto las variables de alto rendimiento, como las variables clínicas.

Se calcularon los intervalos de confianza (IC) para los modelos que mostraron mayor rendimiento, y se realizó un análisis de interpretabilidad del mejor regresor. Sin duda, una de las grandes ventajas que presentan las regresiones de Cox frente a otros modelos de supervivencia es la interpretabilidad del mismo, que permite extraer conclusiones directas acerca de los riesgos aportados por cada covariable. Por último, se comprobó el supuesto de proporcionalidad de riesgos, visualizando los residuales escalados de Schoenfeld, que son el resultado del escalado inverso de los residuales de Schoenfeld en base a su varianza. Los residuales de Schoenfeld se obtienen de la diferencia entre las covarianzas observadas y las esperadas. Gráficos horizontales advierten del cumplimiento del supuesto de proporcionalidad.

3.5.3.3. Cox-Net: regresión de Cox con *elastic net penalty*

Cox-Net se refiere a la adición de regularización *lasso* (l1) o *ridge* (l2) al modelo de Cox, con el objetivo de evitar el *overfitting* e incrementar la generalización del modelo. La cantidad de regularización se parametriza a través de los hiperparámetros α, que determinan la cantidad de regularización a la que se someterá modelo, y r, que determina la proporción de penalización l1.

Por lo tanto, un $\alpha = 0$ daría como resultado una regresión de Cox estándar, así como un $r = 0$ resultaría en una regresión *ridge*, un $r = 1$ en una regresión *lasso*, y un valor intermedio en la combinación de ambas, denominada *penalización neta elástica*, dado lugar a la llamada Cox-Net o regresión de Cox con penalización neta elástica [49].

Al igual que en el caso anterior con la regresión de Cox estándar, se ajustó una Cox-Net mediante el empleo de CV para la identificación de los hiperparámetros óptimos, se realizaron análisis de explicabilidad, se calcularon los IC al 95% por *bootstrapping*, y se realizaron predicciones con el modelo final (*Anexo B*). Cabe mencionar que como se detalla en el *Anexo B*, a pesar de que la regresión de Cox no necesitó de la estandarización de los datos de entrada por tratarse de un modelo lineal, la adición de la regularización para conformar la Cox-Net obligó a realizar este preprocesado.

3.6. *Machine Learning*

Clásicamente la regresión de Cox se ha postulado como el modelo de análisis de supervivencia estándar a nivel global [18,22]. Sin embargo, en las últimas décadas, se han adaptado algunos algoritmos de ML al manejo de datos censurados, lo que ha conllevado la irrupción de esta disciplina de la IA en el campo del análisis de supervivencia [18,22].

3.6.1. Análisis de supervivencia: aproximación basada en ML

La regresión de Cox continúa siendo a día de hoy una de las herramientas predilectas para el estudio de las variables *time-to-event*, ya que presenta importantes ventajas como una alta capacidad predictiva, una gran sencillez computacional, y una elevada interpretabilidad de los resultados [22].

No obstante, este modelo no se encuentra exento de inconvenientes. Entre sus desventajas destacan tres por su gran relevancia: asume proporcionalidad de *hazards*; no es capaz de modelizar adecuadamente relaciones no lineales ni interacciones; y es una técnica diseñada para analizar muestras de pequeño tamaño, por lo que escala con dificultad a datos de alta dimensionalidad [22].

En este contexto se han erigido algunos algoritmos de ML adaptados para el manejo de datos censurados como alternativas a la regresión de Cox. De hecho, ya se pueden encontrar en la literatura autores que defienden el *sorpasso* de algunos modelos de IA al clásico modelo de CPH [22]. Actualmente, los algoritmos de ML que cuentan con mayor nivel de adaptación para el análisis de supervivencia son: el *Survival Support Vector Machine* (SSVM), que surge de la adaptación del *Support Vector Machine* (SVM), el *Random Survival Forest* (RSF) conformado por la unión de varios *Survival Trees* (ST), que son a su vez adaptaciones de los *Decision Trees* (DT), los modelos *Gradient Boosted* (GB), y los AFT, si bien estos últimos no son algoritmos exclusivos de ML.

3.6.1.1. Reducción de la dimensionalidad: PCAs y FA

Se llevaron a cabo dos técnicas de reducción de la dimensionalidad: un análisis de componentes principales (PCA) [50] y un análisis factorial (FA) [51] de datos mixtos. El objetivo de estos análisis fue resumir el *dataset* en un pequeño grupo de variables (componentes o factores) que pudiera servir para ajustar los modelos posteriormente.

En primer lugar se realizó un PCA, una de las técnicas más conocidas de reducción de la dimensionalidad y eliminación de la multicolinealidad. Dado que este tipo de análisis es sensible a la escala, se realizó una estandarización de las variables cuantitativas antes de comenzar.

En segundo lugar se realizó un FA para datos mixtos, también con estandarización de las variables cuantitativas de entrada. A pesar de que existen autores que han apuntado a la utilidad y aplicabilidad de los PCAs en presencia de datos categóricos recogidos en tipos numéricos [18], los PCAs son una técnica diseñada para actuar sobre variables cuantitativas. Por esta razón, se consideró conveniente la realización de un FA teniendo en cuenta que se trata de una técnica diseñada específicamente para manejar *datasets* en los que coexisten variables numéricas y categóricas.

El FA construye una matriz que representa las relaciones entre las variables del conjunto de datos. En el caso de las variables cuantitativas, esta matriz es la matriz de covarianza, mientras que en el caso de las variables categóricas, es la matriz de disimilitudes. Posteriormente, se trata de identificar las direcciones que maximizan la varianza para las variables numéricas y la discriminación para las variables categóricas. Esto quiere decir que el FA aplica una técnica similar al PCA en los datos de naturaleza numérica, y una técnica similar al análisis de correspondencias múltiples en el caso de las categóricas.

Por último, es importante resaltar una diferencia sustancial entre los PCAs y los FAs. Mientras que el objetivo de los PCAs es puramente encontrar unas variables llamadas componentes que consigan explicar la mayor cantidad de varianza por sí mismas, los FAs pretenden encontrar otras variables llamadas factores que sean interpretables y entendibles, a la vez que sean capaces de resumir en sí mismas la máxima información contenida en varias de las variables originales.

3.6.1.2. Métodos *kernel*: SVM y SSVM

Entre los algoritmos que fueron seleccionados para modelizar los datos de esta investigación, se encontraban dos de la familia de los métodos kernel: el SVM y su adaptación al análisis de supervivencia, el SSVM.

Una característica común de estos métodos es su sensibilidad a la escala de los datos de entrada, ya que la distancia entre los datos en el hiperespacio de representación en el que delimitan el hiperplano, puede depender directamente de la magnitud de la escala de los datos. Por esta razón, se comenzó realizando

un preprocesamiento encaminado a eliminar valores perdidos y a estandarizar los datos cuantitativos de entrada.

Es importante destacar también que los SVM no son *a priori* buenos modelos para combinar variables cuantitativas con variables categóricas, sin embargo, si se realiza una estandarización de las variables cuantitativas y las categóricas se encuentran codificadas por números, se puede llegar a alcanzar buenos resultados con estos algoritmos, ya que las distancias euclídeas entre los datos serían similares.

Se entrenaron por un lado tres SVM para predecir tres outcomes: supervivencia al ingreso (dicotómico), curación tras el episodio (dicotómico), y evolución (con tres clases: no deterioro, deterioro, y fallecimiento). Se acopló un PCA en una *pipeline*, pues a menudo mejoran el rendimiento de los SVM. Los hiperparámetros optimizados para estos tres primeros entrenamientos fueron: n_*componentes* del *PCA*, y C, *Kernel* y gamma del SVM. La optimización se realizó a través de una CV y los modelos fueron testados mediante la matriz de confusión y las métricas de precisión diagnóstica. También se estimaron los IC por *bootstrapping*, tal y como se detalla en el *Anexo B*.

A continuación se entrenaron los SSVMs siguiendo las primeras dos de las tres aproximaciones que han sido planteadas a la hora de resolver problemas de supervivencia mediante el empleo de SVM, que son:

- Ranqueo [52]: trata el problema como si de un problema de clasificación ordinal se tratase. El objetivo del algoritmo no es predecir el tiempo de supervivencia, sino la asignación de *score* de riesgo que permita la ordenación de un individuo en una serie. Maneja de forma satisfactoria los datos censurados.
- Regresor [53]: basado en la idea de aplicar una *Support Vector Regression* para resolver el problema tratando la variable *Survival* como una variable dependiente y típica de un problema de regresión. El objetivo en este caso es predecir el logaritmo del tiempo. Presenta el inconveniente de no contar con un buen manejo de los datos censurados, de hecho, cuando se propuso originariamente esta técnica, se llegó a proponer que las censuras debían ser ignoradas.
- Híbrido: combina la predicción del orden con la predicción del tiempo.

El total de SSVM entrenados fue de tres: uno con kernel lineal y objetivo de ranqueo; otro con kernel lineal y objetivo de regresión; y otro con kernel clínico y objetivo de ranqueo. En los tres casos el *outcome* fue supervivencia al ingreso.

Los hiperparámetros a optimizar fueron diferentes a los de los SVM. Los SSVM optimizaron el kernel, el hiperparámetro α, que determina el nivel de regularización, y el hiperparámetro r, que determina el objetivo del SSVM, a saber ranqueo ($r = 1$), regresión ($r = 0$), o híbrido ($0 < r < 1$). La técnica de entrenamiento empleada fue la misma que para los SVM, y se detalla en el *Anexo B*. Posteriormente se calcularon los IC al 95% por *bootstrapping*, pero si bien en el caso anterior de los SVM estándar la métrica de evaluación fue la *accuracy*, en este caso se utilizaron los *c-index* de Harrell.

3.6.1.3. Métodos de *ensemble*: RSF, GB, y AFT

Los métodos de *ensemble* consisten en la asociación de varios modelos independientes que al actuar de forma conjunta mejoran su capacidad predictiva. En otros campos de la IA, estos métodos han demostrado altos niveles de rendimiento. En lo referente al análisis de supervivencia, se han propuesto los RSF, los modelos GB, y los modelos AFT.

Los RSF son una extensión de los bosques aleatorios. Al igual que estos, están formados por el ensamblaje de múltiples árboles entrenados con muestras *bootstrap*. Sin embargo, he aquí la diferencia entre los RSF y los bosques estándar. Mientras que los primeros se componente de ST, los segundos se componen de DT. La diferencia entre un ST y un DT no es más que el criterio de división. Los DT tienden a maximizar la pureza de sus nodos valiéndose de métricas, como Gini o la entropía, mientras que los ST tienden a maximizar el estadístico *logrank*, es decir, tienden a maximizar la diferencia entre supervivencias [18].

Los RSF permiten modelizar datos complejos y no lineales, así como procesar eficientemente datos de alta dimensionalidad, identificar interacciones, e imputar datos perdidos de forma natural [22]. Estas características están propiciando el incrementando de su aparición en publicaciones relativas a análisis de supervivencia en los últimos años, y su cada vez más asentada posición como alternativa de la regresión de Cox.

Para el entrenamiento de los RSF, se comenzó con el análisis de los hiperparámetros *profundidad* y *número de estimadores*, que permitieron guiar la *GridSearch* y reducir el tiempo de computación. Posteriormente, se siguió el mismo proceso de entrenamiento que para el resto de modelos, y se calcularon los IC al 95% mediante *bootstrapping*. Como única particularidad asociada al modelo, se calculó la importancia asociada a cada variable por permutación, y se evaluaron predicciones individuales a varios pacientes.

Los siguientes modelos de *ensemble* entrenados emplearon la estrategia denominada GB, una técnica ML capaz de optimizar una función de pérdida diferenciable mediante la construcción de un modelo basado en la unión de varios submodelos de predicción débil, que individualmente son escasamente capaces de mejorar la aleatoriedad, pero que ensamblados en serie pueden llegar a lograr gran capacidad predictiva [54].

Los modelos ensamblados fueron árboles de decisión y modelos de regresión lineal. La función de pérdida utilizada en ambos casos fue la *likelihood* parcial de la CPH. Se estudió el hiperparámetro *número de estimadores* en ambos modelos antes de definir la *GridSearch* para reducir el tiempo de entrenamiento, y se realizó el mismo proceso que para los RSF.

Por último se ajustó un modelo AFT. Si bien como se ha descrito previamente existen estimadores de la función de supervivencia no paramétricos (*Kaplan-Meier*) y semi-paramétricos (CPH), los AFT representan estimadores paramétricos. Estos modelos difieren de los anteriores en que si bien en el modelo CPH se estudia el efecto de las covariables sobre el riesgo *hazard* mientras el tiempo fluye, en los AFT las covariables afectan directamente al tiempo [55].

En el modelo CPH el efecto de las covariables se obtiene sobre la función de *hazard*. En este caso, si el *hazard basal* se considera paramétrico, se podría modelizar el comportamiento con un modelo de Weibull, exponencial o de Gompertz [55]. Si por el contrario no se considera paramétrico, el modelo obtenido será el CPH. En el modelo AFT en cambio, se analiza el efecto de las covariables sobre el logaritmo del tiempo de supervivencia [55]. Los modelos obtenidos en este caso incluyen gamma generalizado, Log-logístico, Log-normal, Weibull y exponencial [55]. Weibull y exponencial son los únicos modelos de regresión paramétrica que tienen tanto una representación de riesgos proporcionales como de tiempo de fallo acelerado [55].

El modelo AFT se ajustó con función de pérdida de probabilidad inversa de censura ponderada por mínimos cuadrados, y siguiendo el mismo protocolo que en los casos anteriores. Se estudió el hiperparámetro *número de estimadores*, se definió una *GridSearch*, se encontraron los hiperparámetros óptimos mediante CV, se calcularon los IC al 95%, y la importancia de las variables, y se realizaron predicciones.

En lo referente a este último punto, cabe destacar de los métodos de *ensemble* que las predicciones naturales difieren de unos a otros: RSF devuelve las curvas de supervivencia y *hazard*, mientras que GB y AFT devuelven un score.

3.7. Investigación abierta y reproducible, y reutilización de datos.

El material utilizado para llevar a cabo este proyecto se ha publicado en un repositorio abierto de GitHub [31], con la única excepción de algunas bases de datos que contienen atributos sensibles. Sin embargo, se ha comprobado que los *datasets* compartidos preservan toda la información necesaria para reproducir los resultados. De esta forma, también se han compartido los cuadernos de programación ejecutados con los resultados obtenidos.

Una de las principales preocupaciones de este proyecto, además del propio análisis de los datos, ha sido la construcción de una base de datos completa y de calidad que pueda ser utilizada por otros investigadores en nuevas investigaciones. Por esta razón, los *datasets* se han acompañado de ficheros de metadatos que documentan con detalle las características de cada una de las variables. El fin último es permitir la reutilización de los datos, adaptándolos a los principios FAIR.

3.8. Recursos computacionales.

Esta investigación se llevó a cabo utilizando dos lenguajes de programación: *Python* (versión 3.6.8) [56] y *R* (versión 4.1.0) [57]. De entre todas las librerías, módulos, y paquetes utilizados, tuvo especial protagonsimo la librería de Python *scikit-survival* (versión 0.14.1), un módulo construido sobre *scikit-learn* que trata de aprovechar las ventajas de este segundo a la par que implementa nuevas funciones para el tratamiento de series *time-to-event* [49]. El resto de librerías, módulos y paquetes utilizados, pueden consultarse en el repositorio del trabajo [31], donde se encuentra compartido todo el código de programación de este proyecto.

Como terminal de trabajo se utilizó un ordenador personal modelo MacBook Pro de principios del 2015, con 8 GB de RAM, y un procesador de 2.7 GHz Intel Core i5 de doble núcleos.

4. RESULTADOS

3. Resultados

Como se ha descrito previamente, este proyecto llevó a cabo un análisis de más de 130 variables pertenecientes a pacientes con ICH que incluyó descripciones estadísticas, contrastes de hipótesis múltiples, síntesis de gráficos científicos, y modelización de datos con algoritmos estadísticos y de ML. Por razones de legibilidad, interpretabilidad y extensión, en este manuscrito se expondrán y discutirán únicamente los resultados más relevantes, si bien se pueden encontrar todos los resultados arrojados por la investigación en el repositorio abierto del trabajo [31].

4.1. Pacientes reclutados

Se reclutaron un total de 300 pacientes con una edad media de 69 ± 16 años y una mediana de edad que alcanzó 71 ± 21 años, comprendiendo un rango de edad de entre 0 y 101 años (*Tabla C1*). De los 300 pacientes, 118 (39.33%) fueron mujeres y 182 (60.67%) fueron varones (*Tabla C2*). El centro que aportó más pacientes fue la Institución - 1 con un total de 257 (85.67%), seguido de Institución - 2 con 17 (5.67%) pacientes, y de la Institución - 3 con 11 (3.67%); no se contó con el registro de procedencia de los 15 (5%) pacientes (*Tabla C2*). El ingreso más temprano ocurrió el 30-11-2009, y el más tardío el 18-11-2015 (*Tabla C3*). Las distribuciones de edad (*Figura C2*), sexo (*Figura C7)*, hospital de procedencia (*Figura C7)* y fechas de ingreso (*Figura C6)*, se pueden encontrar en el *Anexo C*.

4.2. Estadística descriptiva y EDA

Al final del estudio se comprobó que el número de supervivientes al ingreso fue de 209 (69.67%), mientras que los 91 (30.33%) restantes, fallecieron (*Tabla C2*). La mediana de supervivencia (*survival_days*) del subgrupo de pacientes fallecidos durante el ingreso (*survival_discharge* = 0) fue de 6 ± 9 días (*Tabla C1*).

De los supervivientes, 98 (32.67%) pacientes alcanzaron la curación completa (*final_outcome* = 0), 96 (32.00%) quedaron con secuelas (*final_outcome* = 1), y 4 (1.33%) fallecieron en los 3 meses posteriores (*final_outcome* = 1) (*Tabla C2*). De los 11 restantes no se obtuvieron datos que permitiesen clasificarlos adecuadamente, por lo que corresponden con 11 de los 16 valores perdidos (*final_outcome* = NA) (*Tabla C2*). Los 5 valores perdidos restantes corresponden a pacientes fallecidos en el hospital por causa desconocida, y suman en conjunto con los 82 (27.33%) fallecidos durante el ingreso a causa de la ICH (*final_outcome* = 2) y los 4 (1.33%) fallecidos durante el ingreso por otra causa (*final_outcome* = 3), los 91 no supervivientes al ingreso (*survival_discharge* = 1) (*Tabla C2*).

Del total de hemorragias, 132 (44.00%) fueron primarias y 168 (56%) secundarias. Dentro de las primeras, las más frecuentes fueron las hipertensivas que incluyeron 84 (28%), y dentro de las segundas las más frecuentes fueron traumáticas, con 133 (44.33%) hemorragias (*Tabla C2*).

Recibieron tratamiento neuroquirúrgico (*neurosurg* = 1) 122 (40.67%) pacientes, y a 6 (2.00%) pacientes se les colocó una válvula de derivación (*neurosurg* = 2). Solo 6 (2.00%) pacientes fueron tratados mediante procedimiento intervencionista (*Tabla C2*).

Por último, cabe destacar algunas variables que fueron de gran utilidad para la construcción de los modelos. Entre ellas se encuentran: la glucosa, cuyo nivel medio al ingreso fue de 151.45 ± 51.12 mg/dL; la actividad de protrombina, que mostró una actividad media de 78.08 ± 28.30 %; las horas que transcurrieron desde que comenzó la clínica hasta que el paciente fue a urgencias (*onset_h*), que mostró una media de 3 ± 22 horas; y los días de hospitalización, con una media de 16.93 ± 15.58 días. También resultaron relevantes el número de antecedentes médicos previos de los pacientes (*nfamily_medhist*), que de media resultó en 5 ± 3; la toma de cumarínicos en 58 (19.33%) pacientes; o la presencia de alteración del nivel de conciencia a la llegada a urgencias presente en 159 (53%) pacientes. La puntuación GCS (*tgcs*) fue detectada por varios modelos como una de las variables predictoras principales, siendo sus valores más frecuentes 15, en 77 (25.67%) pacientes, 14 en 49 (16.33%) pacientes, y 3 en 24 (8.00 %) pacientes.
El resto de estadísticos descriptivos puede consultarse en el Anexo C y en el repositorio del trabajo [31].

4.3. Estadística inferencial: contrastes de hipótesis

Previamente a la realización de los contrastes de hipótesis entre las variables del estudio, se comenzó con la comprobación del supuesto de normalidad. Como se indicó previamente en *Materiales y métodos*, el criterio de elección para decidir la normalidad fue la morfología en los gráficos Q-Q. Teniendo en cuenta esta prueba, se consideró que tres variables seguían la normalidad: la presión arterial sistólica, la concentración sanguínea de hemoglobina, y la concentración sérica de potasio (*Figura 3*). De las tres variables, la hemoglobina (p = 0.06) y el potasio (p = 0.06) fueron no significativas en el test de *Shapiro-Wilk*, lo que va a favor del supuesto de normalidad, sin embargo, la presión sistólica fue significativa (p = 0.03) (*Tabla D1*).

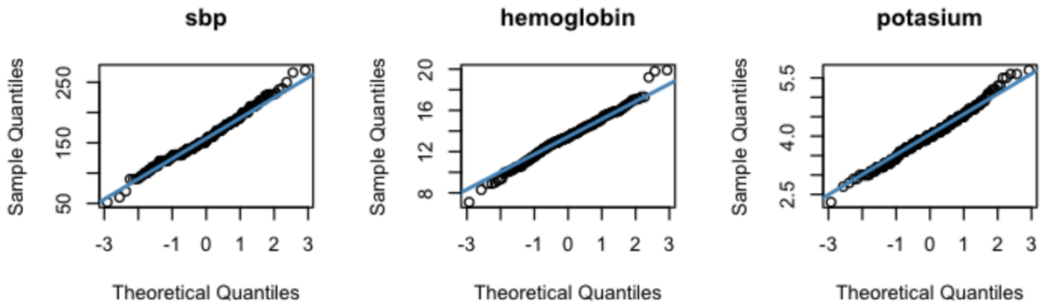


Figura 3. Gráficos Q-Q de las variables que se consideraron normales.

Los análisis de correlación demostraron las correlaciones esperables entre variables que recogían información similar, como la existente entre la hemoglobina y el hematocrito (ρ = 0.99, $p < 0.001$), o entre el filtrado glomerular (*egfr*) y la creatinina (ρ = 0.69, $p < 0.001$) o la urea (ρ = 0.67, $p < 0.001$). Además de éstas, pudieron observarse otras correlaciones interesantes como la existente entre los niveles de fibrinógeno y la actividad de protrombina (ρ = 0.71, $p < 0.001$), los niveles de fibrinógeno y los días de supervivencia globales (ρ = 0.71, $p < 0.001$), y la actividad de protrombina y los días de supervivencia globales (ρ = 0.59, $p < 0.001$) (*Tabla D2*). Resultados posteriores apuntaron a la actividad de protrombina como la variable correlacionada con los niveles de fibrinógeno y con la supervivencia, con la que podría presentar también asociación causal; mientras que la correlación entre los niveles de fibrinógeno y la supervivencia podría ser una correlación espúrea.

Otras correlaciones observadas fueron: días de hospitalización en UCI y niveles de glucosa (ρ = 0.54, $p < 0.001$), inr y supervivencia global (ρ = -0.63, $p < 0.001$), y cantidad de alcohol ingerido (*g_alcohol*) y filtrado glomerular renal (ρ = -0.59, $p < 0.0001$) (*Tabla D2*). Las matrices de correlación pueden consultarse en la *Figura D1*.

En cuanto a los contrastes de hipótesis entre las variables del estudio, únicamente cuatro comparaciones cumplieron los supuestos expuestos en la *Tabla 6* del test de la T de *Student* (*Tabla D3*) (*Figura D2*), y únicamente una los de la U de Mann-Whitney-Wilcoxon (*Tabla D4*). Estos test apuntaron a la existencia de asociaciones estadísticamente entre la concentración de hemoglobina sérica y la supervivencia a los tres meses ($p = 0.021$), un año ($p = 0.020$), cinco años ($p = 0.010$) y al tiempo total de seguimiento ($p = 0.011$).

Los test de la mediana de *Mood* por su parte (*Tabla D5*) (*Figura D3*), mostraron una esperanza de vida global menor en pacientes con alta comorbilidad, especialmente en aquellos que presentaban arritmias ($p = 0.015$), enfermedades cardiovasculares ($p < 0.035$) y enfermedades neurológicas ($p = 0.019$). Otro aspecto

que también pareció condicionar la mediana de supervivencia fue el padecimiento de una ICH primaria (p = 0.047) y la toma de cumarínicos (p = 0.015).

Por otro lado, se asociaron a una menor supervivencia el número de antecedentes familiares en varios tramos, como por ejemplo a los 15 días (p = 0.024), a los 5 años (p = 0.006), o al alta (p = 0.044); los niveles de glucosa también en varios tramos como a los 3 días (p = 0.018), a los 1 mes (p = 0.002), o 5 años (p = 0.048); la actividad de protrombina a los 3 días (p = 0.048), a los 15 días (p = 0.021), o 1 mes (p = 0.004) y 3 meses (p = 0.001); o por ejemplo a los 5 años la edad (p < 0.001), la urea (p = 0.015), los eritrocitos (p = 0.000), o el número de otros fármacos ingeridos (p = 0.034).

En relación al desenlace final, este se vio condicionado por las horas de evolución transcurridas hasta el ingreso (p = 0.010), los niveles de glucosa (p = 0.006), y la actividad de protrombina (p = 0.009), entre otros factores.

Resultó significativa también la asociación entre el número de hospitalizaciones a los 5 años y la presencia de paresias en los miembros derechos, tanto superior (p = 0.047) como inferior (p = 0.047).

Por último, los test χ^2 (*Tabla C7*) (*Figura D4*) mostraron resultados coherentes a las observaciones de los otros test. El desenlace final mostró diferencias entre los que presentaban afectación de los miembros derechos superior (p = 0.030) e inferior (p = 0.019), diferente GCS (p = 0.003), una ICH primaria (p = 0.002), o enfermedades previas como arritmias (p = 0.022), entre otros. La supervivencia también vio condicionada por la presencia de clínica neurológica (p = 0.034), los antecedentes de cardiopatías (p = 0.002), enfermedades neurológicas (p = 0.002) o haber padecido una ICH primaria (p = 0.028).

4.4. Análisis de supervivencia

4.4.1. Estimador de Kaplan-Meier y *logrank*

Se estimaron las curvas de supervivencia al ingreso global (*Figura 4*) y por subgrupos de covariables potencialmente predictoras (*Figura E1*), que fueron comparadas con el test *logrank* (*Tabla E1*), cuyos resultados reforzaron a la presencia de arritmias (p = 0.036), de síntomas neurológicos (p = 0.039), de alteración del nivel de conciencia (p < 0.001), de bajo puntuaje en la escala GCS (p < 0.001) (*Figura 4*) y de etiología primaria de la ICH (p = 0.049) como factores pronósticos de mortalidad. Otros resultados se muestran en la *Tabla E1*.

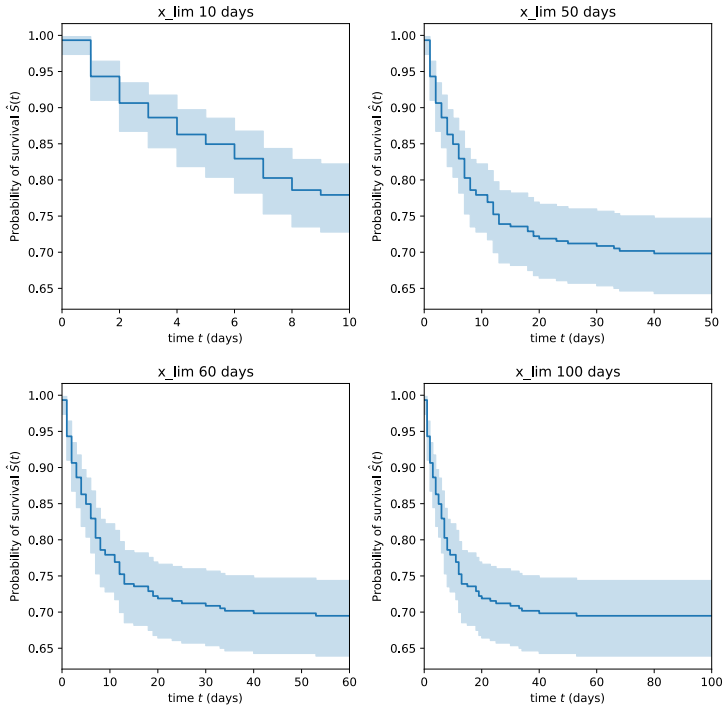


Figura 4. Curvas de supervivencia global estimadas por el *estimador de Kaplan-Meier* con diferentes escalas de tiempo (eje de abscisas)..

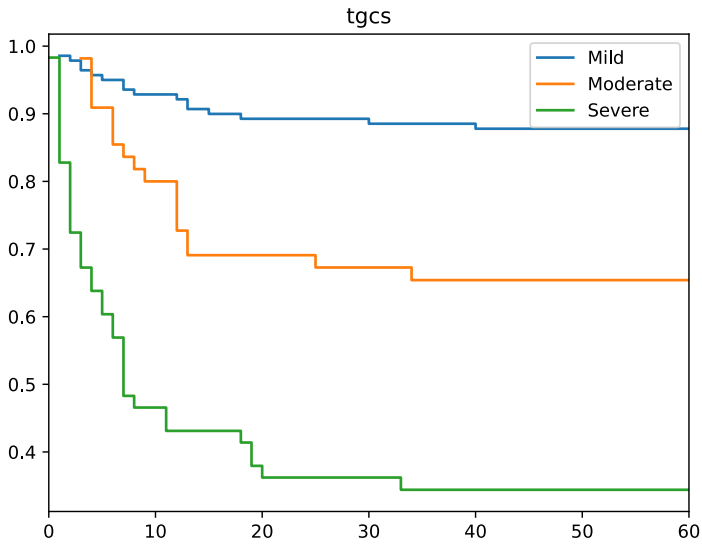


Figura 5. Curvas de supervivencia estimadas por el *estimador de Kaplan-Meier* para tres subgrupos de la variable tgcs (GCS): leve (> 12), moderado (8-12), o grave (<8).

4.4.3. Reducción de la dimensionalidad: PCAs y FA

Los resultados obtenidos por las dos técnicas de reducción de la dimensionalidad empleadas, PCA y FA, demostraron un mayor poder de los FA para resumir la información del *dataset* en menos variables nuevas. Para superar el 85% de varianza explicada, se requirieron 10 componentes vs 4 factores de FA.

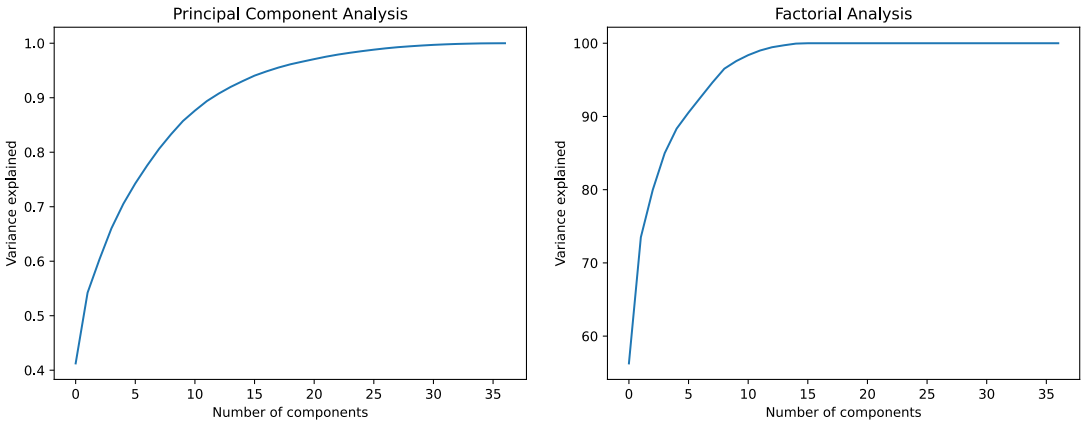


Figura 6. **Resultado de la aplicación de las técnicas de reducción de la dimensionalidad PCA y FA.** Se observan mejores resultados con los factores del FA que con los componentes del PCA.

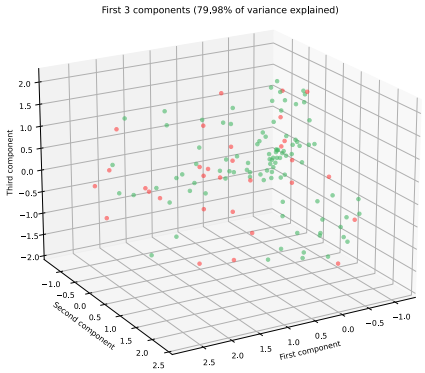


Figura 6. **Ejemplo de la discriminabilidad obtenida con los tres factores principales del FA.** Se observa un cierto grado de separabilidad entre clases de supervivencia al ingreso y una varianza explicada del 80%. En rojo los fallecidos y en verde los supervivientes al ingreso.

4.4.4. Modelos predictivos de supervivencia

Los modelos que obtuvieron mejores rendimientos fueron el CPH y la Cox-Net (*Tabla 7*). Estos dos modelos no solo fueron los que lograron un mayor *c-index*, sino que fueron los modelos con mayor interpretabilidad y que menos características y complejidad computacional presentaron. En concreto, el mayor rendimiento lo obtuvieron CPH1 con un *c-index* de 0.84 IC95% (0.75,0.90), y Cox-Net1 con 0.84 IC95% (0.76,0.90). En segundo lugar, Cox-Net2 con un c-index de 0.77 IC95% (0.69,0.84) y CPH2 con un c-index de 0.76 IC95% (0.67,0.84) se situaron en segundo lugar, utilizando tan solo cinco variables fácilmente extraíbles a la llegada del paciente a urgencias (*Tabla 7*).

Por otro lado, destacó la similitud de los resultados obtenidos por el SVM1 (0.71 IC95% (0.62,0.80)) y su homólogo para supervivencia, el SSVM1 (0.73 (0.62,0.82)). En este caso la modificación del algoritmo para el manejo de datos censurados no incremento el rendimiento de forma significativa. Los modelos GB en su conjunto, obtuvieron los peores resultados (*Tabla 7*).

Tabla 7. Resumen de la complejidad, interpretabilidad y rendimiento de los modelos entrenados

	Variables	*Outcome*	N vars	Pred	Rendimiento**	Interpret.
CPH1	*Tabla E2*	Sp ingreso	6	*S/H(t)*	0.84 (0.75,0.90)	Alta
CPH2	*Tabla E2**	Sp ingreso	5	*S/H(t)*	0.76 (0.67,0.84)	Alta
Cox-Net1	*Tabla E2*	Sp ingreso	6	*S/H(t)*	0.84 (0.76,0.90)	Alta
Cox-Net2	*Tabla E2**	Sp ingreso	5	*S/H(t)*	0.77 (0.69,0.84)	Alta
PCA + SVM1	Tabla E3	Sp ingreso	37	Clase	0.71 (0.62,0.80)	Baja
PCA + SVM2	Tabla E3	*follow_up*	37	Clase	0.49 (0.37,0.61)	Baja
PCA + SVM3	Tabla E3	*final_outcome*	37	Clase	0.73 (0.63,0.83)	Baja
PCA + Linear rank SSVM1	Tabla E3	Sp ingreso	37	*Score*	0.73 (0.62,0.82)	Baja
PCA + Linear regr. SSVM2	Tabla E3	Sp ingreso	37	$Score_t$	0.67 (0.55,0.78)	Baja
PCA + Kernel rank SSVM3	Tabla E3	Sp ingreso	37	*Score*	0.72 (0.61,0.81)	Baja
RSF	Tabla E3	Sp ingreso	37	*S/H(t)*	0.65 (0.53,0.77)	Media
GB *survival trees*	Tabla E3	Sp ingreso	37	*S/H(t)*	0.63 (0.49,0.76)	Media
GB *least squares*	Tabla E3	Sp ingreso	37	*S/H(t)*	0.60 (0.44,0.72)	Media
GB AFT	Tabla E3	Sp ingreso	37	*S/H(t)*	0.66 (0.52,0.78)	Media

Sp = Supervivencia. *Variables de la *Tabla E2* sin *hospitalization_days*. ** Intervalo de confianza al 95% calculado por bootstrapping para la *accuracy* en el caso de los SVM y para el *c-index* en el resto de los modelos.

Otro aspecto fundamental tenido en cuenta además del rendimiento, fue la interpretabilidad de los modelos. Los modelos CPH y Cox-Net mostraron una alta interpretabilidad: CPH permitió incluso extraer conclusiones directas de los propios coeficientes del modelo (*Figura E2*). Los modelos RSF y GB permitieron estratificar la importancia de las características (*Tabla E4*, *Tabla E5* y *Tabla E6*), sin embargo no la cuantificación matemática directa de la fórmula de decisión del algoritmo. SVM y SSVM mostraron la menor interpretabilidad (*Tabla 7).*

Varias de las características que resultaron las más importantes para los modelos, coincidieron con las encontradas como las más significativas en los análisis estadísticos. Por ejemplo, la glucosa, el GCS, la actividad de protrombina o las horas transcurridas entre el inicio de la clínica y la visita a urgencias, se encontraron entre las más importantes para CPH (*Figura E2*), RSF (*Tablas E4*), GB *tree* (*Tabla E5*), y GB AFT (*Tabla E7*). Las características de estos tres modelos fueron similares, mientras que difirieron moderadamente de las encontradas por GB *least squares* (*Tabla E6*).

La interpretabilidad se asoció también a las predicciones dadas por el modelo. Mientras que los SVM, SSVM no permitieron predecir la curva de supervivencia, si no que únicamente daban un *score* de riesgo, CPH, Cox-Net y GB permitieron la predicción de *S(t*) y *H(t)* (*Tabla 7*) (*Figura E3*).

5. DISCUSIÓN Y CONCLUSIONES

5. Discusión y conclusiones

5.1. Identificación de factores pronóstico

Múltiples esfuerzos previos han tratado de identificar factores pronósticos de supervivencia en pacientes con ICH [26]. Muchos de estos factores corresponden con hallazgos en pruebas de imagen [12], y pocos con parámetros extraíbles a través de la anamnesis, exploración física y mediciones sencillas en sangre. Este hecho dificulta la construcción de herramientas diagnósticas para estratificar el riesgo de este tipo de pacientes, las cuales son imprescindibles a la hora de asistir en la toma de decisiones de un planteamiento que puede tener graves consecuencias para la salud y la vida del enfermo.

Tras la recopilación de los datos de 300 pacientes con ICH llevada a cabo en este trabajo, se han logrado identificar algunos parámetros que podrían corresponder con factores de riesgo que podrían ser utilizados para desarrollar escalas clínicas y modelos de riesgo con los que confeccionar futuras guías de práctica clínica.

Dentro de los factores que han resultado más relevantes, destacan los niveles séricos de glucosa, una métrica sencilla, obtenible a través de glucómetros portátiles de bajo coste, que ha presentado una correlación fuerte con los días de ingreso en cuidados intensivos, y asociaciones estadísticamente significativas con la supervivencia y el desenlace final en los test de la mediana. Además, este factor se encontró entre los más importantes para los modelos CPH, RSF, GB *tree* y GB *AFT*. Sin embargo, no es este el primer trabajo que plantea la curiosa relación entre la glucosa y la severidad de una ICH. El metaanálisis llevado a cabo recientemente por *Zheng et al. (2018)*[54] advertía de que esta relación, a pesar de que había sido propuesta por varios autores, no permanecía clara. El citado grupo revisa 16 artículos y llega a la conclusión de que parece existir una asociación entre este parámetro sérico y un pronóstico desfavorable, no obstante advierten de que más estudios son necesarios para establecerla con seguridad. La relación que probablemente explique este comportamiento se encuentra en la acción de la glucosa como osmolito activo, que contribuye al aumento del edema vasogénico asociado al hematoma y a la inducción de la apoptosis neuronal, y a su acción destructiva sobre la barrera hematoencefálica [59].

La actividad de la protrombina también mostró una correlación fuerte con la supervivencia, asociaciones estadísticamente significativas con la supervivencia y el desenlace final, y se encontró entre las principales variables para los modelos CPH, RSF, GB *tree* y GB AFT. Esta asociación ha sido previamente estudiada en la literatura, pero de acuerdo con nuestro conocimiento, siempre en el contexto de pacientes con

tratamiento anticoagulante o antiagregante [60]. Su relación con el pronóstico se atribuye a un mayor aumento del hematoma en las primeras horas de formación a causa de la deficiencia coagulativa manifestada por los niveles séricos bajos de este parámetro. Sin embargo, creemos que puede ser interesante estudiar en el futuro su utilidad en pacientes no anticoagulados, ya que nuestros resultados apuntan a que se podría tratarse de un factor pronóstico independiente.

Como era de esperar, la presencia de clínica neurológica también se postuló como un marcador de pronóstico desfavorable. El test *logrank*, junto con la estratificación de la importancia de los modelos CPH, RSF, GB *least squares*, y GB AFT, señalaron a la alteración del nivel de conciencia como un parámetro asociado a la supervivencia; del mismo modo *logrank*, CPH, RSF, y GB *tree* apuntaron al GCS. Realmente podría considerarse que la variable GCS contiene en sí misma toda la información que contiene la variable alteración del nivel de conciencia, ya que no deja de ser una escala que busca precisamente valorar el deterioro del nivel de vigilia. De nuevo el GCS, es una variable previamente relacionada con el pronóstico [61].

Las horas transcurridas entre el padecimiento de la clínica y la llegada a urgencias resultó en una de las principales características de decisión para CPH, RSF, GB *tree* y GB AFT; y alcanzó la significación estadística en los test de la mediana. De acuerdo con nuestro conocimiento, este factor no ha sido estudiado como tal en la literatura, pero su comportamiento como factor pronóstico parece tener sentido. Los pacientes con clínicas más floridas podrían acudir antes a urgencias que aquellos con clínicas más leves; y como se ha señalado en el párrafo anterior, la clínica en sí guarda estrecha relación con el pronóstico.

Por último, cabe destacar otros dos grupos de factores cuyos resultados sugieren su comportamiento como factores pronóstico en este estudio: la presencia de comorbilidad, como arritmias, enfermedades cardiovasculares, o enfermedades neurológicas previas; y el padecimiento de una ICH primaria con respecto a una secundaria, resultado que se encuentra acorde con la literatura [62].

5.2. Modelos de supervivencia

Los modelos CPH y Cox-Net mostraron mejores resultados que los modelos de ML a la hora de resolver el problema de la predicción de supervivencia. El rendimiento obtenido, con un *c-index* de 0.84 es, de acuerdo con nuestro criterio, un buen resultado que abre la posibilidad al planteamiento de un nuevo estudio que busque validar esta herramienta como asistente en la toma de decisiones en un entorno clínico real. Sin embargo, este resultado es difícil de contrastar con los obtenidos por otros modelos publicados previamente, pues prácticamente en su totalidad incluyen

variables extraídas de la imagen [26], una premisa que este trabajo se propuso suprimir con el objetivo de facilitar su aplicabilidad y extender su uso a regiones geográficas con pocos recursos.

En este sentido, hemos considerado de elección el modelo CPH2, a pesar de que el modelo CPH1 haya presentado mayor rendimiento. La razón de esta elección guarda relación precisamente con la aplicabilidad. Mientras que CPH1 fue entrenado con las variables de la *Tabla E2*, CPH2 fue entrenado con las variables de esta tabla excluyendo los días de hospitalización. Este hecho convierte a CPH2 en un modelo de fácil implantación en un servicio de urgencias, pues únicamente requiere de la introducción de cinco variables que pueden ser recabadas al poco de la llegada del paciente al hospital y en el propio departamento de urgencias: GCS, nivel de glucosa, presencia de alteración del nivel de conciencia, horas desde el inicio de la clínica, y actividad de protrombina.

Otras ventajas de este modelo son su sencillez computacional y su interpretabilidad y *explicabilidad*. Cuatro de las cinco variables de entrada que requiere el modelo son reconocidos factores pronósticos de ICH; y la quinta, las horas de evolución, presenta una explicación plausible y unos resultados en este estudio que apuntan a su comportamiento como factor de riesgo. El hecho de que el modelo se pueda explicar con un criterio clínico es, sin duda, una de las características más importantes del mismo, pues la implantación de herramientas poco interpretables en la práctica médica entraña grandes riesgos [65].

De acuerdo con nuestros resultados, el modelo estadístico CPH ha superado a los modelos de ML a la hora de resolver el análisis de supervivencia. Publicaciones previas han llegado a situar a los modelos de ML por encima de la regresión de Cox, justificando su superioridad en la capacidad de éstos para lidiar con datos multidimensionales, establecer relaciones no lineales y detectar interacciones [18,22]. Sin embargo, nuestro trabajo reafirma la posición de la regresión de Cox como un modelo competitivo y, al menos en algunos ámbitos, al nivel del ML. Probablemente la explicación de estos hallazgos se encuentre en el tamaño muestral. Mientras que los modelos de ML suelen requerir muestras grandes, la regresión de Cox es una solución diseñada para modelizar muestras de datos pequeñas o medianas, como ha ocurrido en nuestro caso [18].

6. REFERENCIAS

6. Referencias

[1]. Caceres JA, Goldstein JN. Intracranial hemorrhage. Emerg Med Clin North Am. 2012 Aug;30(3):771-94. doi: 10.1016/j.emc.2012.06.003.

[2]. Rau CS, Wu SC, Hsu SY, Liu HT, Huang CY, Hsieh TM, Chou SE, Su WT, Liu YW, Hsieh CH. Concurrent Types of Intracranial Hemorrhage are Associated with a Higher Mortality Rate in Adult Patients with Traumatic Subarachnoid Hemorrhage: A Cross-Sectional Retrospective Study. Int J Environ Res Public Health. 2019 Nov 29;16(23):4787.

[3]. Jones HC. Continuity between the ventricular and subarachnoid cerebrospinal fluid in an amphibian, Rana pipiens. Cell Tissue Res. 1978 Dec 14;195(1):153-67.

[4]. Pérez Hernández A, Rodríguez Pérez MDC, Marcelino Rodríguez I, Cuevas Fernández FJ, Domínguez Coello S, Almeida González D, Calleja Puerta S, Cabrera de León A. Incidence and mortality of cerebrovascular disease in Spain: 1,600,000 hospital admissions between 2001 and 2015. Int J Stroke. 2022 Oct;17(9):964-971.

[5]. Rymer MM. Hemorrhagic stroke: intracerebral hemorrhage. Mo Med. 2011 Jan-Feb;108(1):50-4.

[6]. Rajashekar D, Liang JW. Intracerebral Hemorrhage. [Updated 2023 Feb 6]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK553103/

[7]. Pérez K, Novoa AM, Santamariña-Rubio E, Narvaez Y, Arrufat V, Borrell C, Cabeza E, Cirera E, Ferrando J, García-Altés A, Gonzalez-Luque JC, Lizarbe V, Martin-Cantera C, Seguí-Gómez M, Suelves JM; Working Group for Study of Injuries of Spanish Society of Epidemiology. Incidence trends of traumatic spinal cord injury and traumatic brain injury in Spain, 2000-2009. Accid Anal Prev. 2012 May;46:37-44.

[8]. Tenny S, Thorell W. Intracranial Hemorrhage. [Updated 2023 Feb 13]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470242/

[9]. Kanematsu R, Hanakita J, Takahashi T, Park S, Minami M. Radiologic Features and Clinical Course of Chronic Spinal Epidural Hematoma: Report of 4 Cases and Literature Review. World Neurosurg. 2018 Dec;120:82-89.

[10]. Spetzler RF, McDougall CG, Zabramski JM, Albuquerque FC, Hills NK, Nakaji P, Karis JP, Wallace RC. Ten-year analysis of saccular aneurysms in the Barrow Ruptured Aneurysm Trial. J Neurosurg. 2019 Mar 08;132(3):771-776.

[11]. Sahni R, Weinberger J. Management of intracerebral hemorrhage. Vasc Health Risk Manag. 2007;3(5):701-9.

[12]. An SJ, Kim TJ, Yoon BW. Epidemiology, Risk Factors, and Clinical Features of Intracerebral Hemorrhage: An Update. J Stroke. 2017 Jan;19(1):3-10.

[13]. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according

to age, sex, and ethnic origin: a systematic review and meta-analysis. Lancet Neurol. 2010;9:167–176.

[14]. González-Pérez A, Gaist D, Wallander MA, McFeat G, García-Rodríguez LA. Mortality after hemorrhagic stroke: data from general practice (The Health Improvement Network) Neurology. 2013;81:559–565.

[15]. Feigin VL, Lawes CM, Bennett DA, Barker-Collo SL, Parag V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. Lancet Neurol. 2009;8:355–369.

[16]. Fang MC, Go AS, Chang Y, Hylek EM, Henault LE, Jensvold NG, et al. Death and disability from warfarin-associated intracranial and extracranial hemorrhages. Am J Med. 2007;120:700–705.

[17]. Shreffler J, Huecker MR. Survival Analysis. [Updated 2023 May 22]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560604/

[18]. Spooner A, Chen E, Sowmya A, Sachdev P, Kochan NA, Trollor J, Brodaty H. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep. 2020 Nov 23;10(1):20410.

[19]. Rai S, Mishra P, Ghoshal UC. Survival analysis: A primer for the clinician scientists. Indian J Gastroenterol. 2021 Oct;40(5):541-549.

[20]. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003 Jul 21;89(2):232-8. doi: 10.1038/sj.bjc.6601118.

[21]. Definitions: 1. 1. Survival distributions, hazard functions, cumulative hazards [Internet]. Stanford.edu. [cited 2023 Jun 14]. Available from: https://web.stanford.edu/~lutian/coursepdf/unit1.pdf

[22]. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep. 2021 Mar 26;11(1):6968.

[23]. Singh, Jared; Katzman, L. (2018). "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". BMC Medical Research Methodology.

[24]. Nagpal, Chirag (2021). "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks". IEEE Journal of Biomedical and Health Informatics. 25 (8): 3163–3175. arXiv:2003.01176.

[25]. Nagpal, Chirag (2021). "Deep Cox mixtures for survival regression". Machine Learning for Healthcare Conference. arXiv:2101.06536.

[26]. Gregório T, Pipa S, Cavaleiro P, Atanásio G, Albuquerque I, Chaves PC, Azevedo L. Prognostic models for intracerebral hemorrhage: systematic review and meta-analysis. BMC Med Res Methodol. 2018 Nov 20;18(1):145.

[27]. Hemphill JC, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M, et al. Guidelines for the Management of Spontaneous Intracerebral Hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2015;46:2032–60.

[28]. Hwang BY, Appelboom G, Kellner CP, Carpenter AM, Kellner MA, Gigante PR, et al. Clinical grading scales in intracerebral hemorrhage. Neurocrit Care. 2010;13(1):141–51.

[29]. Fritz G, Werner I. Studies on cerebrovascular strokes. II. Clinical findings and short-term prognosis in a stroke mater. Acta Med Scand. 1976;199:133–40.

[30]. Tyrer S, Heyman B. Sampling in epidemiological research: issues, hazards and pitfalls. BJPsych Bull. 2016 Apr;40(2):57-60.

[31]. Menéndez Fernández-Miranda P. Intracranial_Hemorrhages [Internet]. GitHub. [cited 2023 Jun 15]. Available from: https://github.com/Pablomfm/Intracranial_Hemorrhages. doi: 10.5281/zenodo.8040891/.

[32]. Olatunji IE, Rauch J, Katzensteiner M, Khosla M. A Review of Anonymization for Healthcare Data. Big Data. 2022 Mar 10.

[33]. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005 Oct;2(10):e267.

[34]. Rastegar A. Serum Potassium. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Boston: Butterworths; 1990. Chapter 195. Available from: https://www.ncbi.nlm.nih.gov/books/NBK307/

[35]. Fortner B. HDF: The hierarchical data format. Dr Dobb's J Software Tools Prof Program. 1998;23(5):42.

[36]. RDS citation info [Internet]. R-project.org. [cited 2023 Jun 15]. Available from: https://cran.r-project.org/web/packages/RDS/citation.html

[37]. Cooksey RW. Descriptive Statistics for Summarising Data. Illustrating Statistical Procedures: Finding Meaning in Quantitative Data. 2020 May 15:61–139.

[38]. Sullivan GM, Artino AR Jr. Analyzing and interpreting data from likert-type scales. J Grad Med Educ. 2013 Dec;5(4):541-2.

[39]. Guetterman TC. Basics of statistics for primary care research. Fam Med Community Health. 2019 May;7(2):e000067.

[40]. Vickers AJ. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. BMC Med Res Methodol. 2005 Nov 3;5:35.

[41]. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth. 2019 Jan-Mar;22(1):67-72.

[42]. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab. 2012 Spring;10(2):486-9.

[43]. Manuel Gómez-Gómez M, Danglot-Banck C, Vega-Franco L. Sinopsis de pruebas estadísticas no paramétricas. Cuándo usarlas. Revista Mexicana de Pediatría. 2003; 70(2): 91-99.

[44]. Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289-300. Disponible en: http://www.jstor.org/stable/2346101.

[45]. Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, Shukla C, Alm EJ, Hicks SC. A practical guide to methods controlling false discoveries in computational biology. Genome Biol. 2019 Jun 4;20(1):118.

[46]. Altman DG. London (UK): Chapman and Hall; 1992. Analysis of Survival times.In:Practical statistics for Medical research; pp. 365–93.

[47]. Bland JM, Altman DG. The logrank test. BMJ. 2004 May 1;328(7447):1073.

[48]. Abd ElHafeez S, D'Arrigo G, Leonardis D, Fusaro M, Tripepi G, Roumeliotis S. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. Oxid Med Cell Longev. 2021 Nov 30;2021:1302811.

[49]. Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. JMLR. 2020; 21(212):1-6.

[50]. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016 Apr 13;374(2065):20150202.

[51]. van Kesteren EJ, Kievit RA. Exploratory factor analysis with structured residuals for brain network data. Netw Neurosci. 2021 Feb 1;5(1):1-27.

[52]. Van Belle, Pelckmans K, Suykens J, Van Huffel S.Support vector machines for survival analysis. In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED 2007). 2007:1–8.

[53]. Shivaswamy K, Chu W, Jansche M. A support vector approach to censored targets. In Seventh IEEE International Conference on Data Mining (ICDM 2007). 2007: 655-660.Disponible en: https://dx.doi.org/10.1109/ICDM.2007.93.

[54]. Narayanan H, Sokolov M, Butté A, Morbidelli M. Decision Tree-PLS (DT-PLS) algorithm for the development of process: Specific local prediction models. Biotechnol Prog. 2019 Jul;35(4):e2818.

[55]. Zare A, Hosseini M, Mahmoodi M, Mohammad K, Zeraati H, Holakouie Naieni K. A Comparison between Accelerated Failure-time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients. Iran J Public Health. 2015 Aug;44(8):1095-102.

[56]. Van Rossum G, Drake FL Jr, Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam. 1995.

[57]. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2016. Disponible en: https://www.R-project.org.

[58]. Zheng J, Yu Z, Ma L, Guo R, Lin S, You C, Li H. Association Between Blood Glucose and Functional Outcome in Intracerebral Hemorrhage: A Systematic Review and Meta-Analysis. World Neurosurg. 2018 Jun;114:e756-e765.

[59]. Kongwad LI, Hegde A, Menon G, Nair R. Influence of Admission Blood Glucose in Predicting Outcome in Patients With Spontaneous Intracerebral Hematoma. Front Neurol. 2018 Aug 28;9:725.

[60]. Gulati D, Dua D, Torbey MT. Hemostasis in Intracranial Hemorrhage. Front Neurol. 2017 Mar 15;8:80.

[61]. Al-Mufti F, Amuluru K, Lander M, Mathew M, El-Ghanem M, Nuoman R, Park S, Patel V, Singh IP, Gupta G, Gandhi CD. Low Glasgow Coma Score in Traumatic Intracranial Hemorrhage Predicts Development of Cerebral Vasospasm. World Neurosurg. 2018 Dec;120:e68-e71.

[62]. Quiñones-Ossa, G.A., Durango-Espinosa, Y., Padilla-Zambrano, H. *et al.* The puzzle of spontaneous versus traumatic intracranial hemorrhages. *Egypt J Neurosurg* **35**, 13 (2020).

[63]. Lu SC, Swisher CL, Chung C, Jaffray D, Sidey-Gibbons C. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. Front Oncol. 2023 Feb 28;13:1129380.

7. ANEXOS

6. Anexos

8.1. Anexo A: Variables incluidas en el estudio

Tabla A1. Variables incluidas en el estudio tras los procesos de anonimización, limpieza y curado de datos

Variable	Definición	Valores
Auxiliares		
patient	Número de paciente.	De 1 a 300.
time_between_CT_bloodanalysis	Días entre el TC y el análisis de sangre.	De 0 a inf.
Outcomes		
follow_up	Seguimiento clínico.	0 (No deterioro clínico). 1 (Deterioro clínico debido a la ICH). 2 (Fallecimiento debido a la ICH). 3 (Deterioro clínico no debido a la ICH). NA (Valor perdido).
final_outcome	Desenlace final.	0 (Curación completa). 1 (Curación con secuelas). 2 (Fallecimiento hospitalario debido a la ICH). 3 (Fallecimiento hospitalario no debido a la ICH). 4 (Fallecimiento 3 meses post-hospitalización). NA (Valor perdido).
survival_discharge	Supervivencia al ingreso	0 (No); 1 (Sí).
survival_3d	Supervivencia a los 3 días	0 (No); 1 (Sí).
survival_6d	Supervivencia a los 6 días	0 (No); 1 (Sí).
survival_9d	Supervivencia a los 9 días	0 (No); 1 (Sí).
survival_12d	Supervivencia a los 12 días	0 (No); 1 (Sí).
survival_15d	Supervivencia a los 15 días	0 (No); 1 (Sí).
survival_1m	Supervivencia a los 1 mes	0 (No); 1 (Sí).
survival_3m	Supervivencia a los 3 mes	0 (No); 1 (Sí).
survival_1y	Supervivencia a los 1 años	0 (No); 1 (Sí).
survival_5y	Supervivencia a los 5 años	0 (No); 1 (Sí).
survived	Supervivencia al tiempo de seguimiento	0 (No); 1 (Sí).
survival_days	Días de supervivencia	De 0 a inf.
neurosurg		0 (No); 1 (Sí); 2(Válvula DVP)
interprocedures	Tratamiento neurointervencionista	0 (No); 1 (Sí).
Predictors		
sex	Sexo.	1 (Hombre); 2 (Mujer).
hospital	Hospital de procedencia.	1 (Institución-1); 2 (Institución-2); 3 (Institución-3); NA (Valor perdido).
nfamily_medhist	Nº antecedentes familiares.	De 0 a inf.
tobacco	Fumador.	0 (No fumador); 1(Fumador); 2 (Ex-fumador); NA (Valor perdido).
n_tobacco	Paquetes año.	De 0 a inf; NA (Valor perdido).
drugs	Consumo de drogas.	0 (No); 1 (Cocaína); 2 (Cannabis); 3 (Otras); 4 (Ex-consumidor).
alcohol	Consumo de alcohol.	0 (No); 1 (Moderado); 2 (Severo); 3 (Ex-bebedor)

g_alcohol	Gramos de alcohol/día.	De 0 a inf; NA (Valor perdido).
ht	Hipertensión.	0 (No); 1 (Sí).
dmellitus	Diabetes Mellitus.	0 (No); 1 (Sí).
dyslipidemia	Dislipemia.	0 (No); 1 (Sí).
previous_ich	Antecedente de ICH.	0 (No); 1 (Sí).
cv_diseases	Antecedentes cardiovasculares.	0 (No); 1 (Sí).
carrhythmias	Antecedente de arritmias.	0 (No); 1 (Sí).
structural_heart_disease	Antecedente de enfermedad cardiaca.	0 (No); 1 (Sí).
vascular_diseases	Antecedentes de vasculopatía.	0 (No); 1 (Sí).
neurological_diseases	Antecedentes neurológicos.	0 (No); 1 (Sí).
dementia	Antecedente de demencia.	0 (No); 1 (Sí).
depression	Antecedente de depresión.	0 (No); 1 (Sí).
psychiatric_diseases	Antecedentes psquiátricos.	0 (No); 1 (Sí).
cancerous_autoimmune_diseases	Antecedentes autoimmune/oncológico.	0 (No); 1 (Sí).
cancerous_diseases	Antecedentes oncológico.	0 (No); 1 (Sí).
autoimmune_diseases	Antecedentes autoimmune.	0 (No); 1 (Sí).
hyper_hypo_thyroidism	Antecedente hiper/hipotiroidismo.	0 (No); 1 (Sí).
haematological_disorders	Antecedente hemaológicos.	0 (No); 1 (Sí).
other_diseases	Otros antecedentes.	0 (No); 1 (Sí).
antihypertensives	Nº antihipertensivos.	De 0 a inf; NA (Valor perdido).
antidiabetics	Nº antidiabéticos.	De 0 a inf.
hypolipidemics	Nº hipolipemiantes.	De 0 a inf.
anticoagulants	Nº anticoagulantes.	De 0 a inf.
antiplatelets	Nº antiagregantes.	De 0 a inf.
chemotherapeutics	Nº anticoagulants quimioterápicos.	De 0 a inf.
digoxin	Consumo de digoxina.	0 (No); 1 (Sí).
n_other_medications	Nº de otros fármacos.	De 0 a inf.
aceis	Consumo de IECAs.	0 (No); 1 (Sí).
arbs	Consumo de ARA-II.	0 (No); 1 (Sí).
ccbs	Consumo de calcioantagonistas.	0 (No); 1 (Sí).
bblockers	Consumo de beta-bloqueantes.	0 (No); 1 (Sí).
ablockers	Consumo de alfa-bloqueantes.	0 (No); 1 (Sí).
ablockers1	Consumo alfa-bloqueantes anti-HTA.	0 (No); 1 (Sí).
ablockers2	Consumo alfa-bloqueantes no anti-HTA.	0 (No); 1 (Sí).
diuretics	Consumo de diuréticos.	0 (No); 1 (Sí).
other_antihypertensives	Consumo de otros antihipertensivos.	0 (No); 1 (Sí).
biguanides	Consumo de biguanidas.	0 (No); 1 (Sí).
sulfonylureas	Consumo de sulfonilureas.	0 (No); 1 (Sí).
glinides	Consumo de glinidas.	0 (No); 1 (Sí).
glp1a	Consumo de agonistas GLP-1.	0 (No); 1 (Sí).
dpp4i	Consumo de inhibidores de la DPP-4.	0 (No); 1 (Sí).
agi	Consumo de inhibidores de glucosidasa.	0 (No); 1 (Sí).
insulin	Consumo de insulina.	0 (No); 1 (Sí).
statins	Consumo de estatinas.	0 (No); 1 (Sí).
aspirin	Consumo de aspirina.	0 (No); 1 (Sí).

p2y12b	Consumo de bloqueantes P2Y12.	0 (No); 1 (Sí).
gIIbIIIai	Consumo de inhibidores gIIbIIIa.	0 (No); 1 (Sí).
cumarinics	Consumo de cumarínicos.	0 (No); 1 (Sí).
noac	Consumo de NOAC.	0 (No); 1 (Sí).
dabigatran	Consumo de dabigatran.	0 (No); 1 (Sí).
rivaroxaban	Consumo de rivaroxabán.	0 (No); 1 (Sí).
other_medications	Consumo de de otras medicaciones.	0 (No); 1 (Sí).
headache	Cefalea.	0 (No); 1 (Sí); NA (Valor perdido).
emesis	Emesis.	0 (No); 1 (Sí); NA (Valor perdido).
visual_disturbances	Alteraciones visuales.	0 (No); 1 (Sí); NA (Valor perdido).
seizures	Mareos.	0 (No); 1 (Sí); NA (Valor perdido).
mh_trauma	Antecedente de TBI.	0 (No); 1 (Sí); NA (Valor perdido).
mh_le_trauma	Antecedente de TBI de baja energía.	0 (No); 1 (Sí); NA (Valor perdido).
mh_he_trauma	Antecedente de TBI de alta energía.	0 (No); 1 (Sí).
other_symptoms	Otros síntomas.	0 (No); 1 (Sí); NA (Valor perdido).
sbp	PAS (mmHg).	De 0 a inf; NA (Valor perdido).
dbp	PAD (mmHg).	De 0 a inf; NA (Valor perdido).
spo2	SpO2 (%).	De 0 a inf; NA (Valor perdido).
temperature	Temperatura (°C).	De 0 a inf; NA (Valor perdido).
bpm	Pulsaciones por minuto.	De 0 a inf; NA (Valor perdido).
rr	Frecuencia respiratoria.	De 0 a inf; NA (Valor perdido).
neurol_signs	Síntomas neurológicos.	0 (No); 1 (Sí); NA (Valor perdido).
diplopia	Diplopia.	0 (No); 1 (Sí); NA (Valor perdido).
anisocoria	Anisocoria.	0 (No); 1 (Sí); NA (Valor perdido).
aphasia	Aphasia o dysphasia.	0 (No); 1 (Sí); NA (Valor perdido).
dysarthria	Dysarthria.	0 (No); 1 (Sí); NA (Valor perdido).
altered_consciousness	Alteración del nivel de conciencia.	0 (No); 1 (Sí); NA (Valor perdido).
nuchal_rigidity	Rigidez nucal.	0 (No); 1 (Sí); NA (Valor perdido).
rfacial_palsy	Parálisis facial central derecha.	0 (No); 1 (Sí); NA (Valor perdido).
lfacial_palsy	Parálisis facial central izquierda.	0 (No); 1 (Sí); NA (Valor perdido).
ruplimb_mimpairment	Paresia miembro superior derecho.	0 (No); 1 (Sí); NA (Valor perdido).
luplimb_mimpairment	Paresia miembro superior izquierdo.	0 (No); 1 (Sí); NA (Valor perdido).
rlwlimb_mimpairment	Paresia miembro inferior derecho.	0 (No); 1 (Sí); NA (Valor perdido).
llwlimb_mimpairment	Paresia miembro inferior izquierdo.	0 (No); 1 (Sí); NA (Valor perdido).
balance_impairment	Alteración del equilibrio/coordinación.	0 (No); 1 (Sí); NA (Valor perdido).
tgcs	Escala Coma de Glasgow.	De 3 a 15; NA (Valor perdido).
onset_h	Horas desde el inicio de la clínica.	De 0 a inf; NA (Valor perdido).
hospitalizations_1y	Hospitalizaciones en 1 año tras el alta.	De 0 a inf.
hospitalizations_3y	Hospitalizaciones en 3 años tras el alta.	De 0 a inf.
hospitalizations_5y	Hospitalizaciones en 5 años tras el alta.	De 0 a inf.
hospitalization_days	Días de hospitalización.	De 0 a inf.
hospitalization_icu_days	Días de hospitalización en la UCI.	De 0 a inf.
primary_ich	ICH primaria.	0 (No); 1 (Sí).
vascular_ich	ICH de origen vascular.	0 (No); 1 (Sí).
traumatic_ich	ICH traumática.	0 (No); 1 (Sí).

ht_ich	ICH probablemente hipertensiva.	0 (No); 1 (Sí).
amyloidangiopathy_ich	ICH secundaria a angiopatía amiloide.	0 (No); 1 (Sí).
aneurysmal_ich	ICH aneurismática.	0 (No); 1 (Sí).
avm_ich	ICH secundaria a MAV.	0 (No); 1 (Sí).
hti_ich	Transformación hemorrágica (TH) ictus.	0 (No); 1 (TH sin intervencionismo); 2 (TH + intervencionismo no por ictus); 3 (TH + fibrinolisis por ictus); 4 (TH + trombectomía por ictus); 5 (TH + fibrinolisis + trombectomía).
other_ich	ICH secundaria a otras causas.	0 (No); 1 (Sí); NA (Valor perdido).
glucose	Nivel de glucosa (mg/dL).	De 0 a inf; NA (Valor perdido).
urea	Nivel de urea (mg/dL).	De 0 a inf; NA (Valor perdido).
creatinine	Nivel de creatinina (mg/dL).	De 0 a inf; NA (Valor perdido).
sodium	Nivel de sodio (mg/dL).	De 0 a inf; NA (Valor perdido).
potasium	Nivel de potassium (mEq/L).	De 0 a inf; NA (Valor perdido).
egfr	Filtrado glomerular (mL/min/1.73m2)	De 0 a inf; NA (Valor perdido).
prothrombin_activity	Actividad de protrombina (%).	De 0 a inf; NA (Valor perdido).
leukocytes	Leucocitos (10^3/uL).	De 0 a inf; NA (Valor perdido).
erythrocytes	Eritrocitos (10^6/uL).	De 0 a inf; NA (Valor perdido).
hemoglobin	Hemoglobina (g/dL).	De 0 a inf; NA (Valor perdido).
hematocrit	Hematocrito (%).	De 0 a inf; NA (Valor perdido).
platelets	Plaquetas (10^3/uL).	De 0 a inf; NA (Valor perdido).
mcv	Volumen corpuscular medio (fL).	De 0 a inf; NA (Valor perdido).
rdw	ADE.	De 0 a inf; NA (Valor perdido).
mchc	CHCM (g/dL)	De 0 a inf; NA (Valor perdido).
mpv	VPM (fL).	De 0 a inf; NA (Valor perdido).
mch	Hemoglobina corpuscular media (pg).	De 0 a inf; NA (Valor perdido).
inr	INR	De 0 a inf; NA (Valor perdido).
fibrinogen	Nivel de fibrinógeno a la llegada (mg/dL).	De 0 a inf; NA (Valor perdido).
maxfibrinogen	Nivel máximo de fibrinógeno (mg/dL).	De 0 a inf; NA (Valor perdido).
age	Edad (years)	De 0 a inf.

Información más detallada acerca de las variables puede encontrarse en el repositorio del trabajo [31].

8.2. Anexo B: Metodología de entrenamiento, validación y test de los modelos.

Previamente a comenzar con el proceso de entrenamiento de los modelos, se comprobó la existencia de valores perdidos, y en caso de identificarse, se optó por la exclusión completa del sujeto. A pesar de que otras alternativas como la imputación múltiple han sido defendidas en la literatura, ninguna está exenta de críticas. Por esta razón, se decidió un abordaje más conservador que no restara credibilidad a los resultados obtenidos. A continuación, se realizó una estandarización de las variables numéricas en los casos en los que se emplearon algoritmos sensibles a cambios de escala. Este fue el caso de la Cox-Net, los PCAs, el FAs, los SVMs y el SSVM (*Tabla 1A*).

Posteriormente se realizó una partición del dataset en un conjunto de entramiento y otro de test. Estos datos fueron utilizados en un primer momento para realizar pre-entrenamientos dirigidos al estudio de los cambios en el rendimiento del modelo en función de los hiperparámetros más relevantes. Este análisis permitió reducir el espacio de búsqueda y orientar la *GridSearch* del entrenamiento definitivo, lo que conllevó una disminución significativa de los tiempos de entrenamiento, especialmente en el caso de los modelos computacionalmente más complejos, como los modelos GB. Los pre-entrenamientos se realizaron con los conjuntos de entrenamiento y las valoraciones del rendimiento con los conjuntos de test.

Más tarde se definió una *GridSearch* con diferentes combinaciones de hiperparámetros entre los que idealmente se encontraría la configuración óptima, y se llevó a cabo una *K-fold* CV con k = 10 y todo el conjunto de datos como set de entrenamiento (*Tabla A1*). Una vez finalizada la validación cruzada, se reentrenó el modelo final con los hiperparámetros óptimos encontrados.

Por último, se procedió a testar el modelo final. En algunos casos se realizaron análisis de interpretabilidad, y en todos los casos se realizaron predicciones de pacientes del conjunto de test, motivo por el cual se preservó este conjunto no volviendo a utilizar todo el dataset para entrenar el modelo final.

Los test de los modelos concluyeron con la estimación de los IC al 95%, para lo que se utilizó un *bootstrapping* con 10.000 iteraciones. La métrica utilizada para evaluar el rendimiento de los modelos de supervivencia fue, tanto durante el entrenamiento como durante el test, el *c-index* de Harrell. Esta métrica fue diseñada para evaluar modelos de supervivencia en los que no pueden ser empleadas ni las métricas de evaluación de la regresión, como el error cuadrático medio, ni las métricas de evaluación de la clasificación, como la *accuracy*, debido a la existencia de datos censurados. En estos casos el *c-index* provee una solución que estima la proporción de pares concordantes dividida por el número total de pares de evaluación posibles,

es decir, evalúa si el modelo ha rankeado correctamente a una serie de sujetos sin importar la magnitud del riesgo predicho, simplemente importa el orden. Como todo índice, el valor uno implica un rankeo perfecto. En caso de datos con un alto índice de censura, la adaptación del *c-index* descrita por Uno resulta una métrica adecuada que fue utilizada en algunos casos.

Se dieron pequeñas variaciones de la metodología por características particulares de cada algoritmo. Estas variantes se recogen en la *Tabla B1*.

Tabla B1. Especificaciones del entrenamiento y test de los modelos

Modelo	Preprocesamiento	Datos de ajuste	CV-*Folds*	Test	Métrica
CPH	NaN	Todo el conjunto	10	Boots.	*c-index*
Cox-Net	NaN + STD	Todo el conjunto	10	Boots	*c-index*
PCA	NaN + STD	Todo el conjunto	-	-	Varianza
FA	NaN + STD	Todo el conjunto	-	-	Varianza
SVM	NaN + STD	Todo el conjunto	5	Boots + MC	*c-index*
SSVM	NaN + STD	Set de entrenamiento	100	Boots	*Accuracy*
RSF	NaN	Todo el conjunto	10	Boots	*c-index*
Modelo GB	NaN	Todo el conjunto	10	Boots	*c-index*
AFT	NaN	Todo el conjunto	10	Boots	*c-index*

NaN = Valores perdidos; STD = Estandarización; Boots. = *bootstrapping*; MC = matriz de confusión.

8.3. Anexo C: Resultados de estadística descriptiva y EDA

8.3.1. Anexo C1: Variables cuantitativas: tablas y gráficos

Tabla C1. Estadísticos descriptivos de las variables cuantitativas.

Variable	Media	DE	Mediana	IQR	Min	Max	Asimetría	Curtosis	N
Outcome									
survival_days* (días)	8,19	9,30	6.00	9.0	0.00	53.00	2.36	9.59	91,00
Predictors									
nfamily_medhist	5.03	3.30	4.50	4.00	0.00	19.00	1.01	4.58	300.0
n_tobacco (paquetes/año)	8.92	21.86	0.00	0.00	0.00	150.00	3.25	15.99	228.0
g_alcohol (g)	8.33	26.64	0.00	0.00	0.00	200.00	3.79	19.15	252.0
antihypertensives	0.94	1.12	1.00	2.00	0.00	4.00	0.93	2.80	297.0
antidiabetics	0.21	0.53	0.00	0.00	0.00	3.00	2.74	10.61	300.0
hypolipidemics	0.26	0.46	0.00	1.00	0.00	2.00	1.37	3.63	300.0
anticoagulants	0.21	0.41	0.00	0.00	0.00	1.00	1.45	3.10	300.0
antiplatelets	0.29	0.52	0.00	1.00	0.00	2.00	1.59	4.61	300.0
chemotherapeutics	0.01	0.08	0.00	0.00	0.00	1.00	12.12	148.01	300.0
n_other_medications	1.58	1.95	1.00	3.00	0.00	11.00	1.62	6.02	300.0
sbp (mmHg)	158.41	35.25	154.00	45.25	52.00	270.00	0.27	3.29	280.0
dbp (mmHg)	86.39	18.84	85.00	24.25	10.00	145.00	0.12	4.10	280.0
spo2 (%)	96.58	6.12	98.00	5.00	31.00	110.00	-7.07	72.77	189.0
temperatura (°C)	35.83	0.61	36.00	0.50	34.40	38.00	0.34	4.62	67.0
bpm (latidos/min)	80.16	18.11	80.00	24.00	45.00	150.00	0.73	3.76	256.0
rr (respiraciones/min)	16.63	3.67	16.00	4.00	12.00	30.00	1.21	4.60	107.0
onset_h (horas)	47.06	135.47	3.00	22.00	0.00	1464.00	6.75	63.25	205.0
hospitalizations_1y	0.32	0.72	0.00	0.00	0.00	4.00	2.59	9.95	300.0
hospitalizations_3y	0.67	1.15	0.00	1.00	0.00	6.00	2.05	7.39	300.0
hospitalizations_5y	0.92	1.51	0.00	2.00	0.00	8.00	2.20	8.51	300.0
hospitalization_days (días)	17.00	16.00	11.00	16.00	1.00	113	2.29	10.59	300.0
hospitalization_icu_days (días)	6.00	9.00	2.00	8.00	0.00	64.00	2.97	16.12	300.0
glucosa (mg/dL)	151.45	51.52	140.00	57.00	70.00	381.00	1.34	5.20	298.0
urea (mg/dL)	44.01	27.27	39.00	19.00	8.00	278.00	3.86	26.57	298.0
creatinine (mg/dL)	0.91	0.53	0.80	0.30	0.04	6.10	5.51	47.42	298.0
sodium (mg/dL)	139.21	4.73	140.00	5.00	109.00	152.00	-1.97	11.59	297.0
potasium (mEq/dL)	4.02	0.56	4.00	0.70	2.30	5.70	0.22	3.31	290.0
egfr (mL/min/1.73m^2)	57.44	8.04	60.10	0.00	8.00	60.10	-3.53	16.14	275.0
prothrombin_activity (%)	78.08	28.30	90.00	29.00	6.00	118.00	-1.23	3.15	293.0
leukocytes (por 10^3 µL)	10.60	6.11	9.35	4.78	3.00	67.40	4.97	41.74	298.0
Erythrocytes (por 10^6 µL)	4.37	0.67	4.40	0.70	2.40	8.90	0.90	9.76	298.0
hemoglobin (mg/dL)	13.43	1.89	13.40	2.30	7.10	19.90	-0.02	3.76	298.0

hematocrit (%)	39.82	6.13	39.95	6.88	4.90	61.10	-0.75	7.21	298.0
platelets (por 10^3 µL)	201.45	74.27	196.50	76.75	17.00	750.00	2.12	15.28	298.0
mcv (fL)	91.87	7.65	92.00	6.00	8.80	118.00	-4.20	49.64	293.0
rdw (%)	14.36	1.95	14.00	1.50	3.60	28.80	2.07	18.36	289.0
mchc (g/dL)	33.36	2.29	33.50	1.10	3.10	38.20	-9.66	119.42	293.0
mpv (fL)	9.50	9.24	8.40	1.50	3.60	123.00	10.16	110.77	285.0
mch (pg)	30.88	2.44	30.90	2.60	19.30	39.90	-0.13	6.41	293.0
inr	1.54	1.31	1.08	0.25	0.83	12.01	3.99	23.74	282.0
fibrinogen (mg/dL)	410.07	125.56	394.00	163,00	69.00	967.00	0.86	4.95	245.0
maxfibrinogen (mg/dL)	609.78	220.24	560.00	315,00	69.00	1319.00	0.63	2.94	271.0
age (años)	69.00	16.00	71.00	21.00	0.00	101.00	-1.10	4.80	300.0

DE = desviación estándar; IQR = Rango intercuartílico. *Estadísticos descriptivos de la variable días de supervivencia (*survival_days)* para el subgrupo de pacientes que no sobrevivieron al ingreso (*survival_discharge* = 0). No se presentan los estadísticos descriptos globales de la variable, ni tampoco los del subgrupo supervivientes al ingreso, dado que a efectos del análisis de supervivencia posterior, se consideran datos censurados a la derecha.

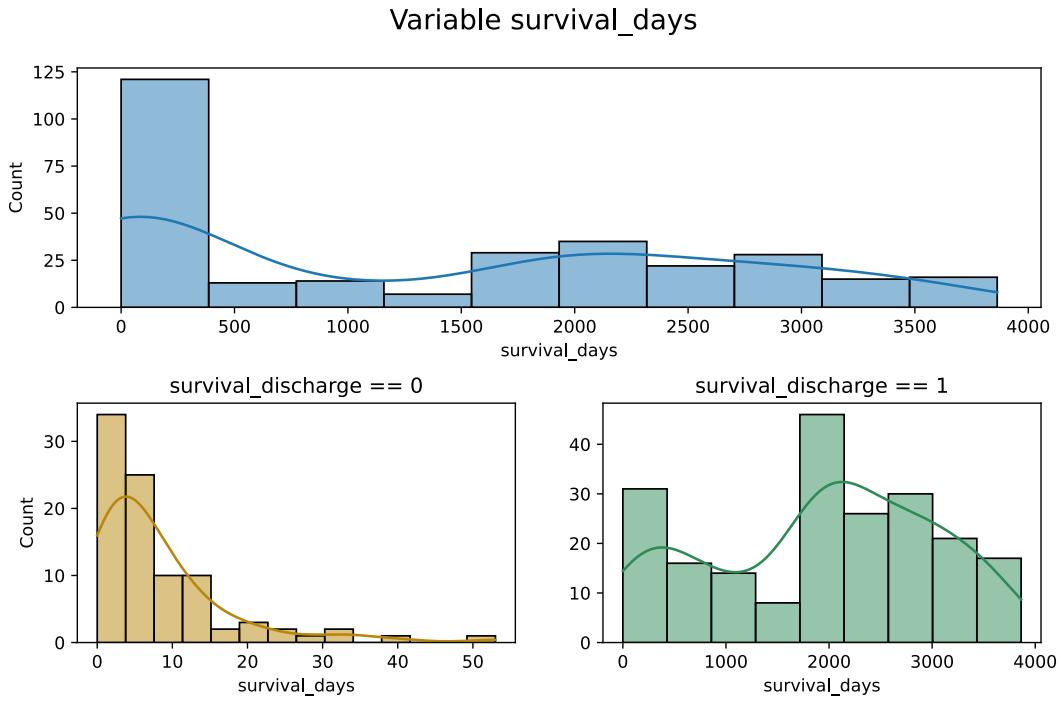


Figura C1. **Distribución de la variable días de supervivencia (*survival_days*).** Histograma de la distribución de la variable *survival_days*, incluyendo la distribución en los subgrupos de no supervivientes al ingreso (*survival_discharge* = 0) y de supervivientes al ingreso (*survival_discharge* = 0). Se pueden consultar más gráficos al respecto en el repositorio del trabajo [31].

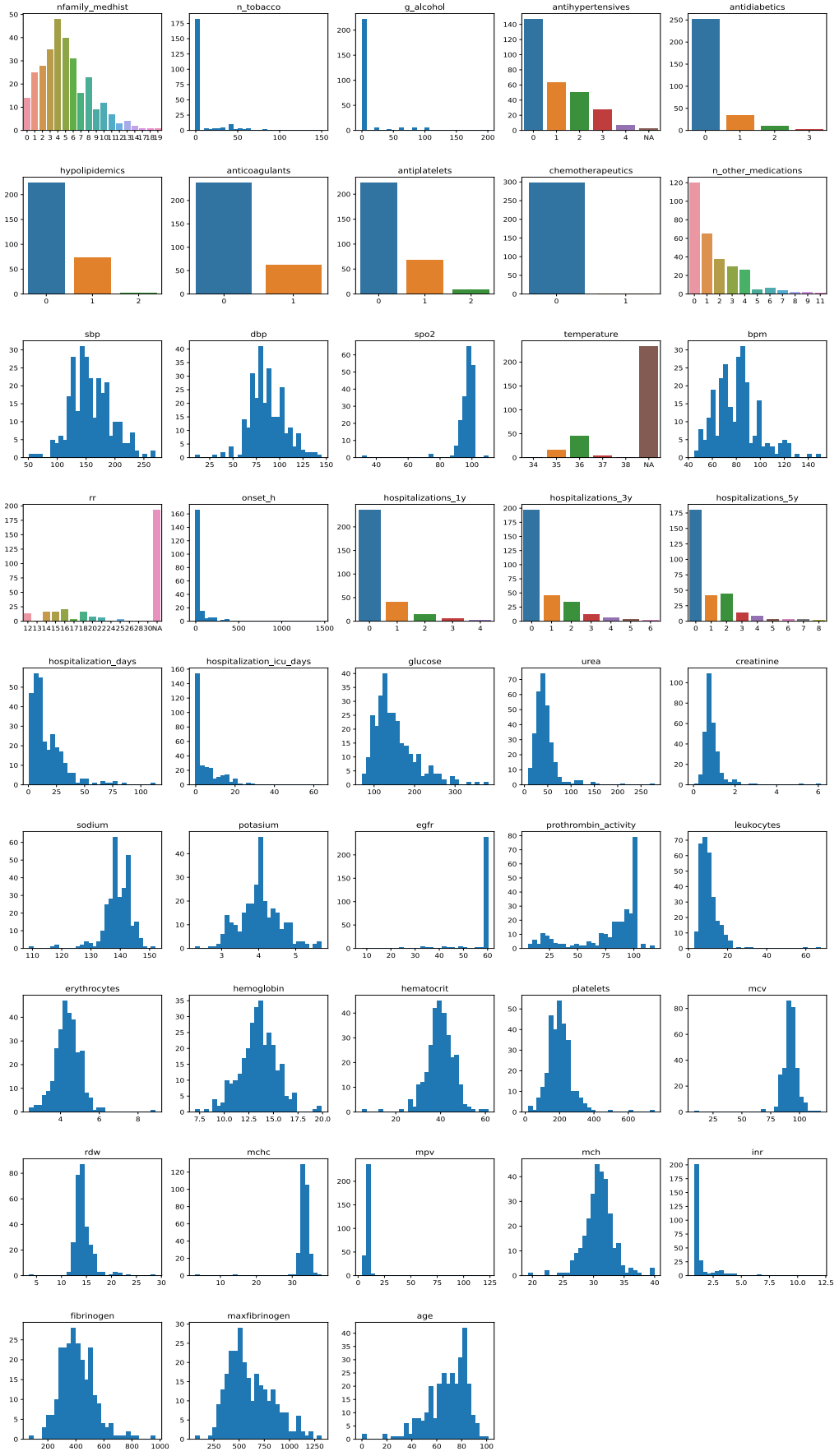


Figura C2. **Distribución de las variables cuantitativas predictoras.** Diagramas de barras para las variables discretas e histogramas para las variables continuas. Se pueden consultar más gráficos al respecto en el repositorio del trabajo [31].

8.3.2. Anexo C2: Variables cualitativas: tablas y gráficos

Tabla C2. Estadísticos descriptivos de las principales variables cualitativas.

Variable	n (%)
Outcomes	
follow_up	
0 (No deterioro clínico)	133 (44.33)
1 (Deterioro clínico debido a la ICH)	66 (22.00)
2 (Fallecimiento debido a la ICH)	93 (31.00)
3 (Deterioro clínico no debido a la ICH)	6 (2.00)
NA (Valor perdido)	2 (0.67)
final_outcome	
0 (Curación completa)	98 (32.67)
1 (Curación con secuelas)	96 (32.00)
2 (Fallecimiento hospitalario debido a la ICH)	82 (27.33)
3 (Fallecimiento hospitalario no debido a la ICH)	4 (1.33)
4 (Fallecimiento 3 meses post-hospitalización)	4 (1.33)
NA (Valor perdido)	16 (5.33)
survival_discharge	
0 (No)	91 (30.33)
1 (Sí)	209 (69.67)
survival_3d	
0 (No)	26 (8.67)
1 (Sí)	274 (91.33)
survival_6d	
0 (No)	45 (15.00)
1 (Sí)	255 (85.00)
survival_9d	
0 (No)	65 (21.67)
1 (Sí)	235 (78.33)
survival_12d	
0 (No)	71 (23.67)
1 (Sí)	229 (76.33)
survival_15d	
0 (No)	80 (26.67)
1 (Sí)	220 (73.33)
survival_1m	
0 (No)	94 (31.33)
1 (Sí)	206 (68.67)
survival_3m	
0 (No)	99 (33.00)

1 (Sí)	201 (67.00)
survival_1y	
0 (No)	116 (38.67)
1 (Sí)	184 (61.33)
survival_5y	
0 (No)	163 (54.33)
1 (Sí)	137 (45.67)
survived	
0 (No)	183 (61.00)
1 (Sí)	117 (39.00)
neurosurg	
0 (No)	172 (57.33)
1 (Sí)	122 (40.67)
2(Válvula DVP)	6 (2.00)
interprocedures	
0 (No)	294 (98.00)
1 (Sí)	6 (2.00)
Predictores	
sex	
1 (Hombre)	182 (60.67)
2 (Mujer)	118 (39.33)
hospital	
1 (Institución - 1)	257 (85.67)
2 (Institución - 2)	17 (5.67)
3 (Institución - 3)	11 (3.67)
NA (Valor perdido)	15 (5.00)
ht	
0 (No)	129 (43.00)
1 (Sí)	171 (57.00)
previous_ich	
0 (No)	287 (95.67)
1 (Sí)	13 (4.33)
depression	
0 (No)	274 (91.33)
1 (Sí)	26 (8.67)
dementia	
0 (No)	266 (88.67)
1 (Sí)	34 (11.33)
digoxin	
0 (No)	275 (91.67)
1 (Sí)	25 (8.33)
aspirin	227 (75.67)
0 (No)	73 (24.33)

1 (Sí)	
cumarinics	
0 (No)	242 (80.67)
1 (Sí)	58 (19.33)
headache	
0 (No)	224 (74.67)
1 (Sí)	74 (24.67)
NA (Valor perdido)	2 (0.67)
emesis	
0 (No)	239 (79.67)
1 (Sí)	59 (19.67)
NA (Valor perdido)	2 (0.67)
visual_disturbances	
0 (No)	285 (95.00)
1 (Sí)	13 (4.33)
NA (Valor perdido)	2 (0.67)
seizures	
0 (No)	282 (94.00)
1 (Sí)	15 (5.00)
NA (Valor perdido)	3 (1.00)
other_symptoms	
0 (No)	137 (45.67)
1 (Sí)	161 (53.67)
NA (Valor perdido)	2 (0.67)
neurol_signs	
0 (No)	43 (14.33)
1 (Sí)	256 (85.33)
NA (Valor perdido)	1 (0.33)
diplopia	
0 (No)	266 (88.67)
1 (Sí)	29 (9.67)
NA (Valor perdido)	5 (1.67)
anisocoria	
0 (No)	261 (87.00)
1 (Sí)	37 (12.33)
NA (Valor perdido)	2 (1.67)
aphasia	
0 (No)	212 (70.67)
1 (Sí)	70 (23.33)
NA (Valor perdido)	18 (6.00)
dysarthria	
0 (No)	237 (79.00)
1 (Sí)	45 (15.00)
NA (Valor perdido)	18 (6.00)

altered_consciousness	
0 (No)	138 (46.00)
1 (Sí)	159 (53.00)
NA (Valor perdido)	3 (1.00)
nuchal_rigidity	
0 (No)	293 (97.67)
1 (Sí)	4 (1.33)
NA (Valor perdido)	3 (1.00)
rfacial_palsy	
0 (No)	265 (88.33)
1 (Sí)	25 (8.33)
NA (Valor perdido)	10 (3.33)
lfacial_palsy	
0 (No)	261 (87.00)
1 (Sí)	29 (9.67)
NA (Valor perdido)	10 (3.33)
ruplimb_mimpairment	
0 (No)	219 (73.00)
1 (Sí)	59 (19.67)
NA (Valor perdido)	22 (7.33)
luplimb_mimpairment	
0 (No)	222 (74.00)
1 (Sí)	56 (18.67)
NA (Valor perdido)	22 (7.33)
rlwlimb_mimpairment	
0 (No)	218 (72.67)
1 (Sí)	60 (20.00)
NA (Valor perdido)	22 (7.33)
llwlimb_mimpairment	
0 (No)	223 (74.33)
1 (Sí)	55 (18.33)
NA (Valor perdido)	22 (7.33)
balance_impairment	
0 (No)	221 (73.67)
1 (Sí)	38 (12.67)
NA (Valor perdido)	41 (13.67)
tgcs	
3	24 (8.00)
4	8 (2.67)
5	10 (3.33)
6	7 (2.33)
7	10 (3.33)
8	9 (3.00)
9	8 (2.67)

10	14 (4.67)
11	11 (3.67)
12	13 (4.33)
13	14 (4.67)
14	49 (16.33)
15	77 (25.67)
NA (Valor perdido)	46 (15.33)
primary_ich	
0 (No)	168 (56.00)
1 (Sí)	132 (44.00)
vascular_ich	
0 (No)	279 (93.00)
1 (Sí)	21 (7.00)
traumatic_ich	
0 (No)	167 (55.67)
1 (Sí)	133 (44.33)
ht_ich	
0 (No)	216 (72.00)
1 (Sí)	84 (28.00)
amyloidangiopathy_ich	
0 (No)	295 (98.33)
1 (Sí)	5 (1.67)
aneurysmal_ich	
0 (No)	294 (98.00)
1 (Sí)	6 (2.00)
avm_ich	
0 (No)	286 (95.33)
1 (Sí)	14 (4.67)
hti_ich	
0 (No)	287 (95.67)
1 (TH sin intervencionismo)	1 (0.33)
2 (TH + intervencionismo no por ictus)	4 (1.33)
3 (TH + fibrinolisis por ictus)	5 (1.67)
4 (TH + trombectomía por ictus)	2 (0.67)
5 (TH + fibrinolisis + trombectomía)	1 (0.33)
other_ich	
0 (No)	296 (98.67)
1 (Sí)	3 (1.00)
NA (Valor perdido)	1 (0.33)
neurosurg	
0 (No)	172 (57.33)
1 (Sí)	122 (40.67)
2 (Válvula DVP)	6 (2.00)
interprocedures	

0 (No)	294 (98.00)
1 (Sí)	6 (2.00)

TH = Transformación hemorrágica. En esta tabla no se incluyen los estadísticos descriptivos de todas las variables cualitativas, sino de las que han resultado ser las principales para el estudio.

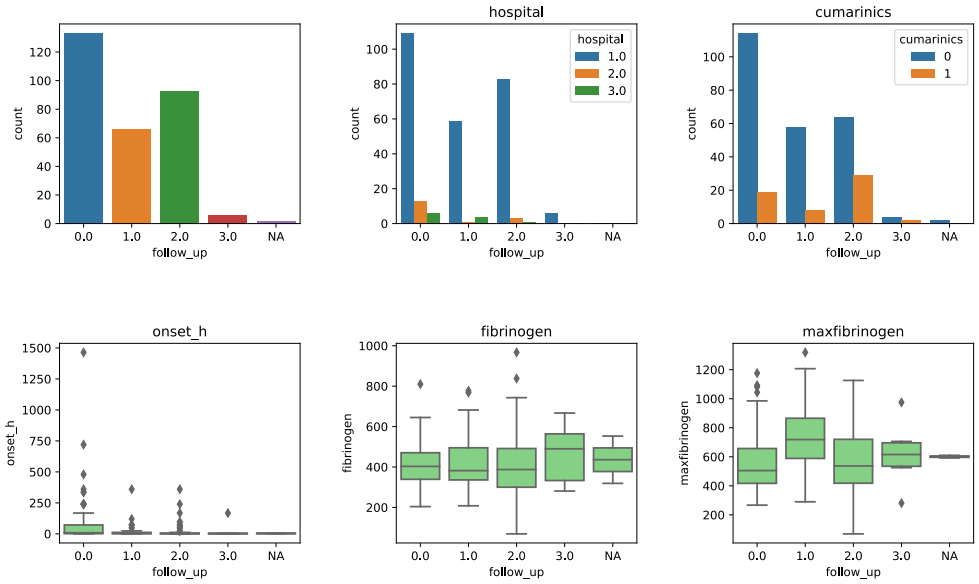


Figura C3. **Distribución de la variable seguimiento (*follow_up*).** Se presenta la distribución de la variable seguimiento de forma global y por subgrupos de variables con interés clínico. Se pueden consultar más gráficos al respecto en el repositorio del trabajo [31].

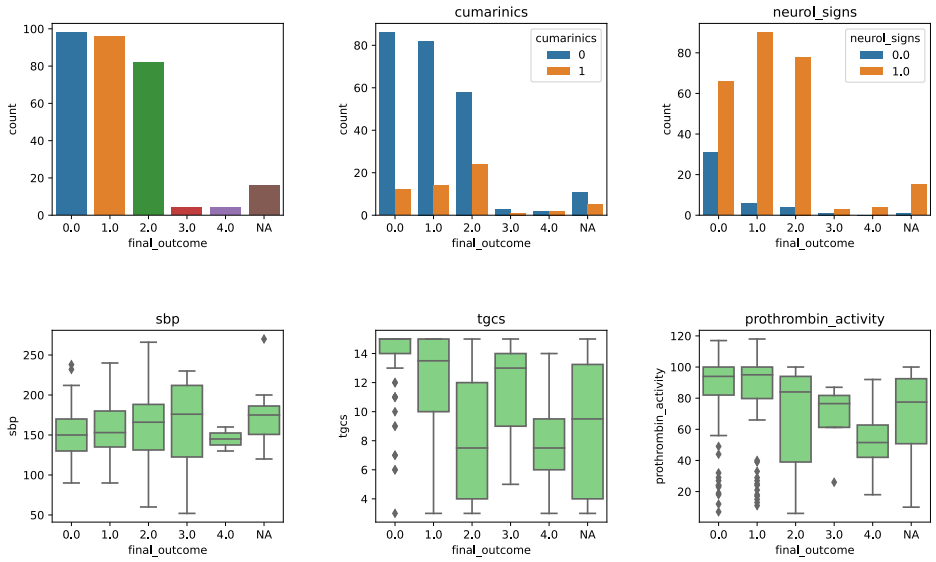


Figura C4. **Distribución de la variable desenlace final (*final_outcome*).** Se presenta la distribución de la variable desenlace final de forma global y por subgrupos de variables con interés clínico. Se pueden consultar más gráficos al respecto en el repositorio del trabajo [31].

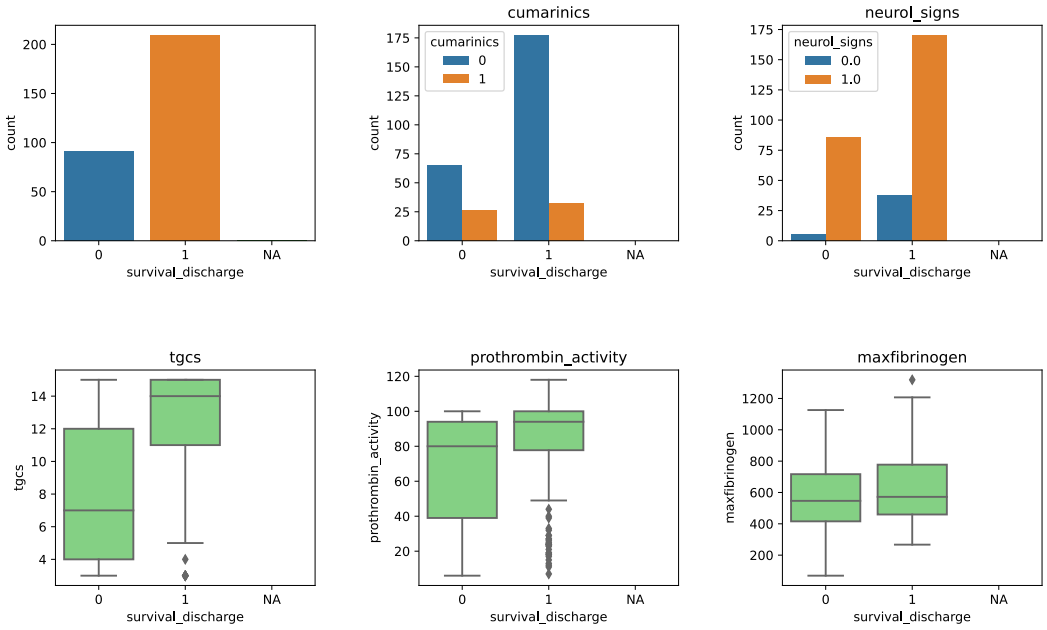


Figura C5. **Distribución de la variable supervivencia al ingreso (*survival_discharge*).** Se presenta la distribución de la variable supervivencia al ingreso, variable que ha dado lugar en el análisis de supervivencia a la variable *Status*. Se presenta su distribución global y por subgrupos de variables con interés clínico. Se pueden consultar más gráficos al respecto en el repositorio del trabajo [31].

Tabla C3. Estadísticos descriptivos de las variables fecha

Variable	Min	Max
Fecha de ingreso	30-11-2009	18-11-2015
Fecha de alta	13-01-2010	27-11-2015
Fecha de TC	30-11-2009	18-11-2015
Fecha de análisis de sangre	30-11-2009	18-11-2015
Fecha de mortalidad	26-01-2010	12-02-2020
Fecha de nacimiento	14-11-1913	07-12-2010

Estas variables no figuran en el dataset porque han sido anonimizadas para preservar la privacidad.

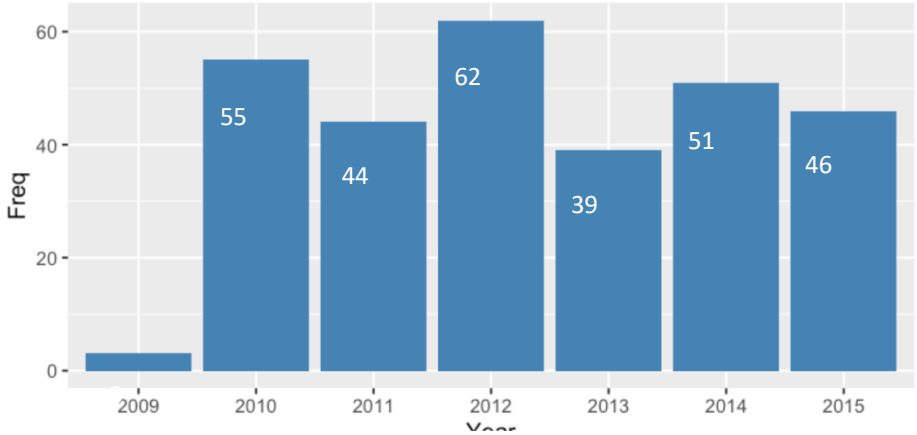


Figura C6. **Distribución de los años de las fechas de ingreso de los pacientes incluidos en el estudio.**

Figura C7. **Distribución de las variables cualitativas predictoras.** Se pueden consultar más gráficos al respecto en el repositorio del trabajo [31].

8.4. Anexo D: Resultados de estadística inferencial: contrastes de hipótesis

Tabla D1. Resultados de los test de Shapiro-Wilk de las variables consideradas normales.

Variable	W	*p*
potasium	0.9907	0.06382
hemoglobin	0.9901	0.06306
sbp	0.9891	0.03351

Se presentan exclusivamente los resultados de las variables consideradas normales, en aras de preservar la legibilidad.

Tabla D2. Resultados más relevantes de los análisis de correlación.

			IC 95%		
Variable 1	Variable 2	*r*/ ρ	Inferior	Superior	*p* ajustado*
fibrinogen	survival_days	0.7099	0.5516	0.8146	<0.0001
prothrombin_activity	fibrinogen	0.7067	0.5768	0.8099	<0.0001
nfamily_medhist	rdw	0.6785	0.528	0.7877	<0.0001
prothrombin_activity	mpv	0.6773	0.5297	0.7853	<0.0001
g_alcohol	leukocytes	0.6644	0.4956	0.7849	<0.0001
sbp	erythrocytes	0.6615	0.5018	0.7776	<0.0001
bpm	platelets	0.6554	0.4854	0.7776	<0.0001
n_other_medications	rdw	0.6548	0.4965	0.771	<0.0001
hospitalization_days	survival_days	0.6347	0.193	0.7566	<0.0001
antidiabetics	rdw	0.6097	0.4736	0.7547	<0.0001
hospitalization_days	fibrinogen	0.5984	0.4059	0.74	<0.0001
hospitalization_days	age	0.5969	0.4377	0.7386	<0.0001
prothrombin_activity	survival_days	0.5893	0.4171	0.7235	<0.0001
hypolipidemics	prothrombin_activity	0.5835	0.4115	0.7195	<0.0001
antihypertensives	rdw	0.5621	0.3924	0.7095	<0.0001
hospitalization_icu_days	glucose	0.5356	0.318	0.6907	<0.0001
g_alcohol	mch	0.525	0.3135	0.6869	<0.0001
g_alcohol	mcv	0.5082	0.3274	0.67	<0.0001
g_alcohol	egfr	-0.5949	-0.7178	-0.4073	<0.0001
inr	survival_days	-0.6337	-0.7572	-0.4664	<0.0001

*Los valores de p han sido corregidos con la técnica para test múltiples FDR [44]. Se presentan únicamente los resultados más relevantes. El resto de los resultados se puede consultar en el repositorio abierto del trabajo [31].

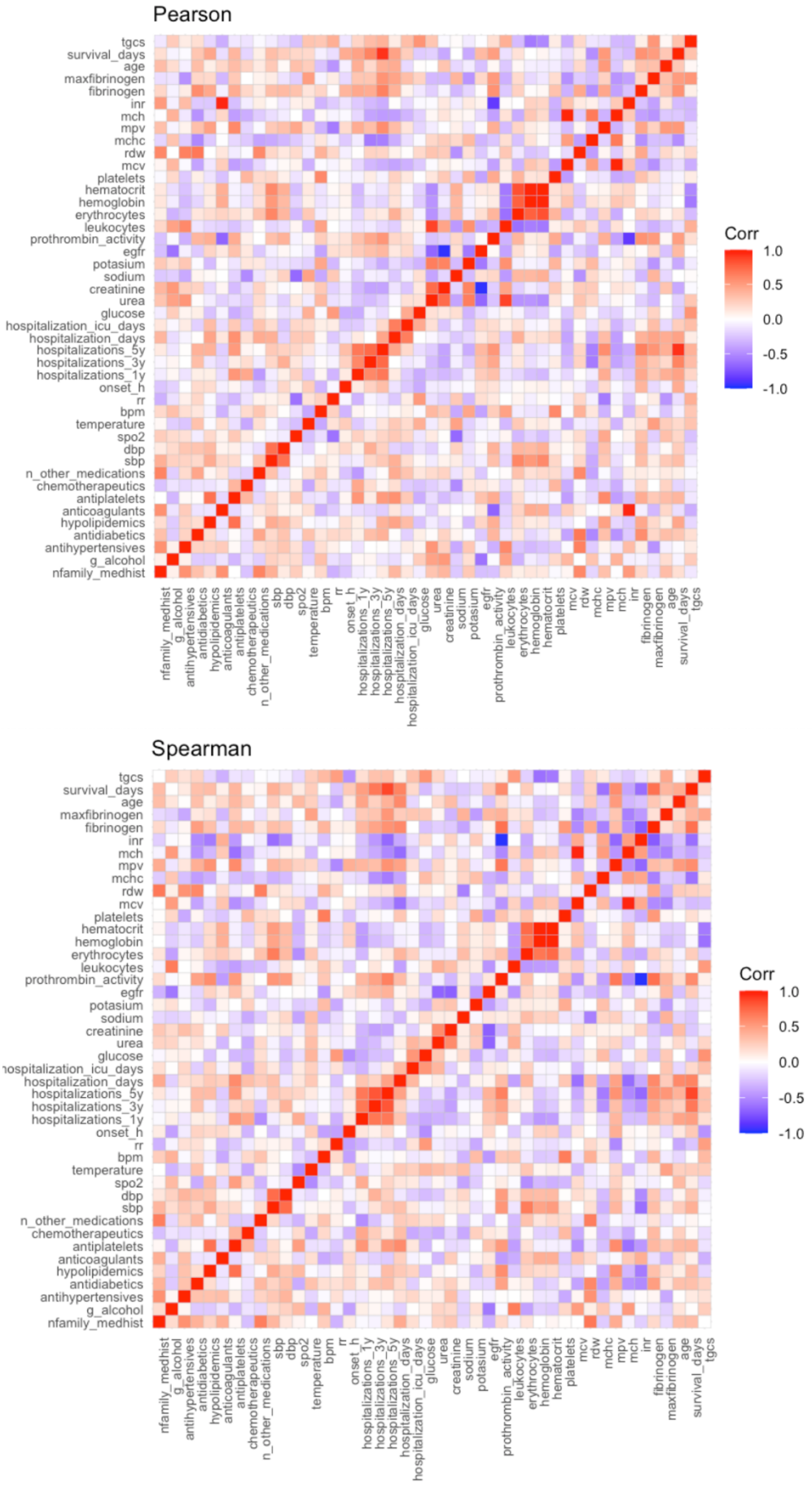


Figura D1. Matrices de correlación de *Pearson* y *Spearman*.

Tabla D3. Comparaciones realizadas mediante el test T de *Student*.

Variable 1	Variable 2	*p* ajustado*
survival_3m	hemoglobin	0.0264
survival_1y	hemoglobin	0.0195
survival_5y	hemoglobin	0.01
survived	hemoglobin	0.0109

*Los valores de *p* han sido ajustados con la técnica para test múltiples FDR [44]. Se presentan las muestras que han sido comparadas con el test T de *Student*, por cumplir los supuestos requeridos para la realización del mismo, detallados en la *Tabla 6*.

Tabla D4. Comparaciones realizadas mediante el test de la U de *Mann-Whitney-Wilcoxon*.

Variable 1	Variable 2	*p*
survived	fibrinogen	0.1152

Se presenta la única comparación realizada mediante el test de la U de *Mann-Whitney-Wilcoxon*, por ser único que cumplió los supuestos para la realización del mismo, detallados en la *Tabla 6*.

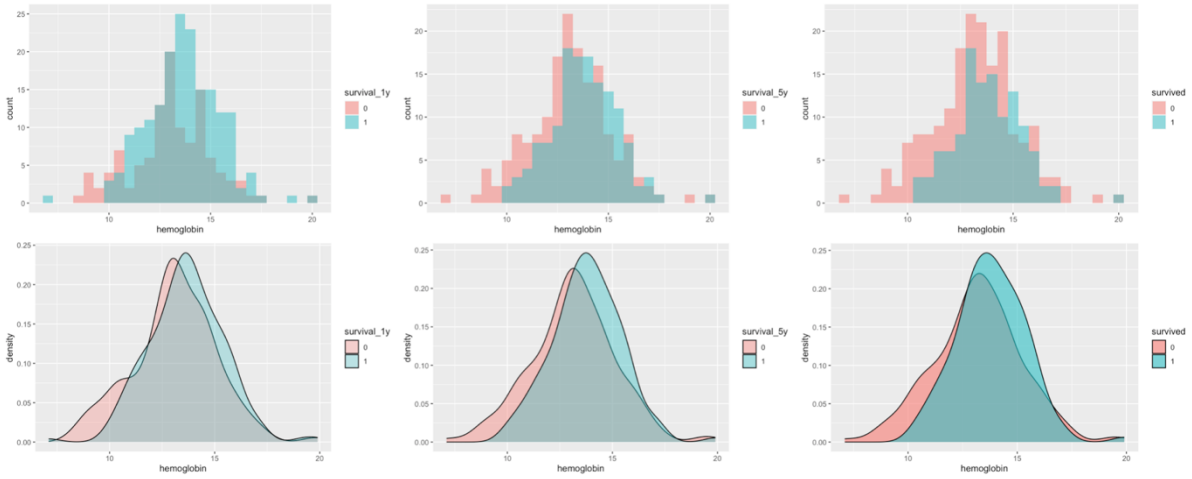
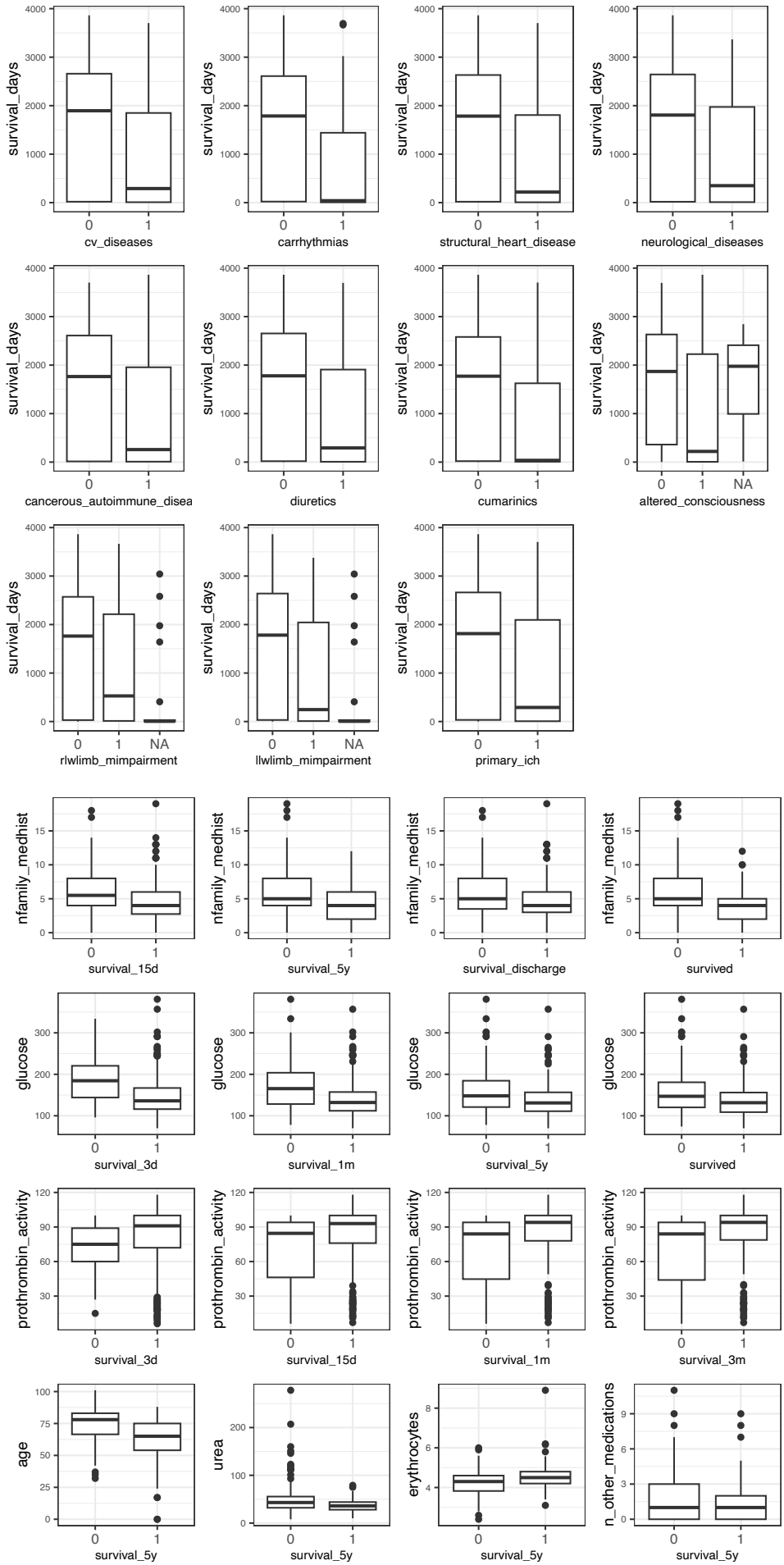


Figura D2. **Distribución de las muestras significativas al test de la T de *Student*.**

Tabla D5. Resultados significativos más relevantes de los test de la mediana.

Variable 1	Variable 2	*p* ajustado*
hospitalizations_5y	ruplimb_mimpairment	0.047
hospitalizations_5y	rlwlimb_mimpairment	0.047
survival_days	cv_diseases	0.0035
survival_days	carrhythmias	0.015
survival_days	structural_heart_disease	0.0035
survival_days	neurological_diseases	0.019
survival_days	cancerous_autoimmune_diseases	0.0466
survival_days	diuretics	0.047
survival_days	cumarinics	0.015
survival_days	altered_consciousness	0.019
survival_days	rlwlimb_mimpairment	0.047
survival_days	llwlimb_mimpairment	0.047
survival_days	primary_ich	0.047
survival_5y	age	< 0.0001
survival_5y	erythrocytes	0.0338
survival_1m	glucose	0.0017
survival_3d	glucose	0.0177
survival_5y	glucose	0.0477
survived	glucose	0.0495
survived	hematocrit	0.0251
survival_discharge	hospitalization_days	< 0.0001
survival_5y	n_other_medications	0.0177
survival_15d	nfamily_medhist	0.024
survival_5y	nfamily_medhist	0.006
survival_discharge	nfamily_medhist	0.0437
survived	nfamily_medhist	0.0021
survival_15d	prothrombin_activity	0.0209
survival_1m	prothrombin_activity	0.0039
survival_3d	prothrombin_activity	0.0477
survival_3m	prothrombin_activity	0.0013
survival_5y	urea	0.015
final_outcome	onset_h	0.0096
final_outcome	glucose	0.0063
final_outcome	prothrombin_activity	0.0094
final_outcome	inr	0.0077
follow_up	glucose	0.0096
follow_up	prothrombin_activity	0.0019
follow_up	maxfibrinogen	<0.0001
neurosurg	nfamily_medhist	0.0152
neurosurg	maxfibrinogen	<0.0001

*Los valores de *p* han sido ajustados con la técnica para test múltiples FDR [44]. Se presentan únicamente los resultados significativos que se han considerado más relevantes. El resto de los resultados se puede consultar en el repositorio abierto del trabajo [31].


survival_days
cv_diseases
carrhythmias
structural_heart_disease
neurological_diseases
cancerous_autoimmune_disea
diuretics
cumarinics
altered_consciousness
rlwlimb_mimpairment
llwlimb_mimpairment
primary_ich
nfamily_medhist
survival_15d
survival_5y
survival_discharge
survived
glucose
survival_3d
survival_1m
survival_5y
survived
prothrombin_activity
survival_3d
survival_15d
survival_1m
survival_3m
age
urea
erythrocytes
n_other_medications
survival_5y

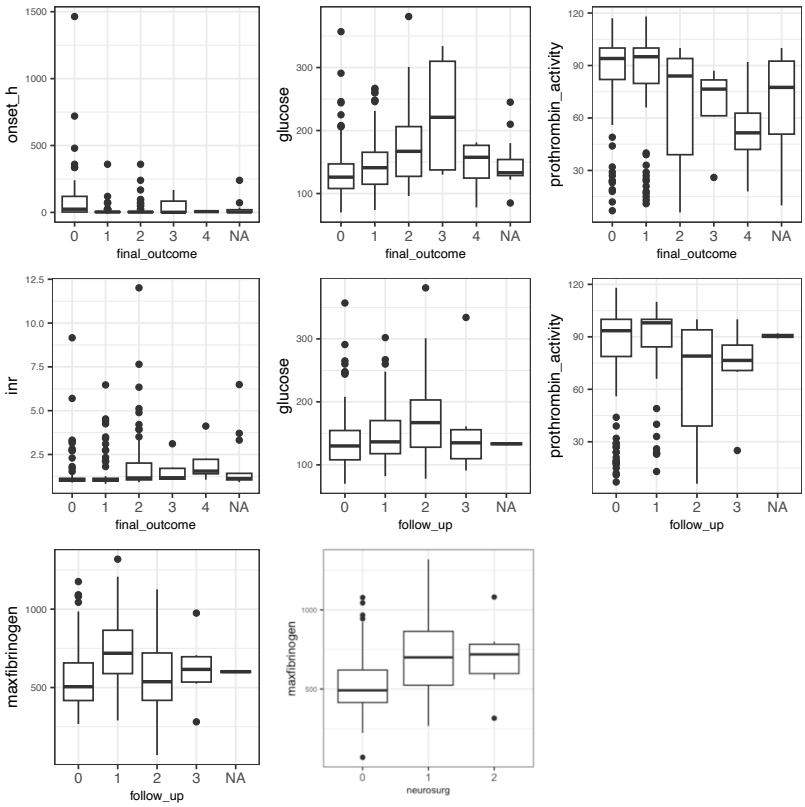


Figura D3. Gráficos de cajas de las variables con las asociaciones estadísticamente significativas más relevantes en los test de la mediana.

Tabla D6. Contrastes *post-hoc* más relevantes tras el test de la mediana.

Grupo 1	Grupo 2	Predictor	*p* ajustado*
final_outcome = 0	final_outcome = 1	onset_h	< 0.0001
final_outcome = 0	final_outcome = 1 2 4	onset_h	< 0.0001
final_outcome = 0 1	final_outcome = 2 4	onset_h	0.0269
follow_up = 0	follow_up = 1 2	onset_h	0.0013
follow_up = 0 1	follow_up = 2	onset_h	0.014
follow_up = 0	follow_up = 1 2	glucose	0.0037
neurosurg = 0	neurosurg = 1	maxfibrinogen	< 0.0001

*Los valores de *p* han sido ajustados con la técnica para test múltiples FDR [44]. Se presentan únicamente los resultados significativos más relevantes. El resto de los resultados se puede consultar en el repositorio [31].

Tabla D7. Resultados significativos al test de χ^2 más relevantes.

Variable 1	Variable 2	*p* ajustado*
final_outcome	dysarthria	0.0372
final_outcome	ruplimb_mimpairment	0.0296
final_outcome	luplimb_mimpairment	0.0098
final_outcome	rlwlimb_mimpairment	0.0192
final_outcome	llwlimb_mimpairment	0.0144
final_outcome	tgcs	0.0029
final_outcome	primary_ich	0.0021
final_outcome	ht_ich	0.0099
follow_up	carrhythmias	0.0216
follow_up	diuretics	0.0415
follow_up	cumarinics	0.0342
follow_up	altered_consciousness	0.0121
follow_up	ruplimb_mimpairment	0.0085
follow_up	luplimb_mimpairment	0.0079
follow_up	rlwlimb_mimpairment	0.0065
follow_up	llwlimb_mimpairment	0.0096
follow_up	balance_impairment	0.0216
interprocedures	ht	0.0415
interprocedures	nuchal_rigidity	0.001
interprocedures	aneurysmal_ich	0.0017
interprocedures	avm_ich	0.015
neurosurg	tgcs	0.0213
neurosurg	primary_ich	0.0112
neurosurg	vascular_ich	0.0399
neurosurg	traumatic_ich	0.0029
neurosurg	aneurysmal_ich	0.0369
survival_discharge	carrhythmias	0.0415
survival_discharge	arbs	0.0192
survival_discharge	neurol_signs	0.0344
survival_discharge	anisocoria	0.0019
survival_discharge	aphasia	0.0022
survival_discharge	balance_impairment	0.0073
survival_discharge	primary_ich	0.028
survival_discharge	traumatic_ich	0.0387
survived	ht	0.0235
survived	structural_heart_disease	0.0017
survived	neurological_diseases	0.0017
survived	dementia	0.0073
survived	cancerous_autoimmune_diseases	0.0018
survived	cancerous_diseases	0.0125
survived	other_diseases	0.0046
survived	cumarinics	0.0024

*Los valores de *p* han sido ajustados con la técnica para test múltiples FDR [44]. Se presentan únicamente los resultados significativos más relevantes. El resto de los resultados se puede consultar en el repositorio [31].

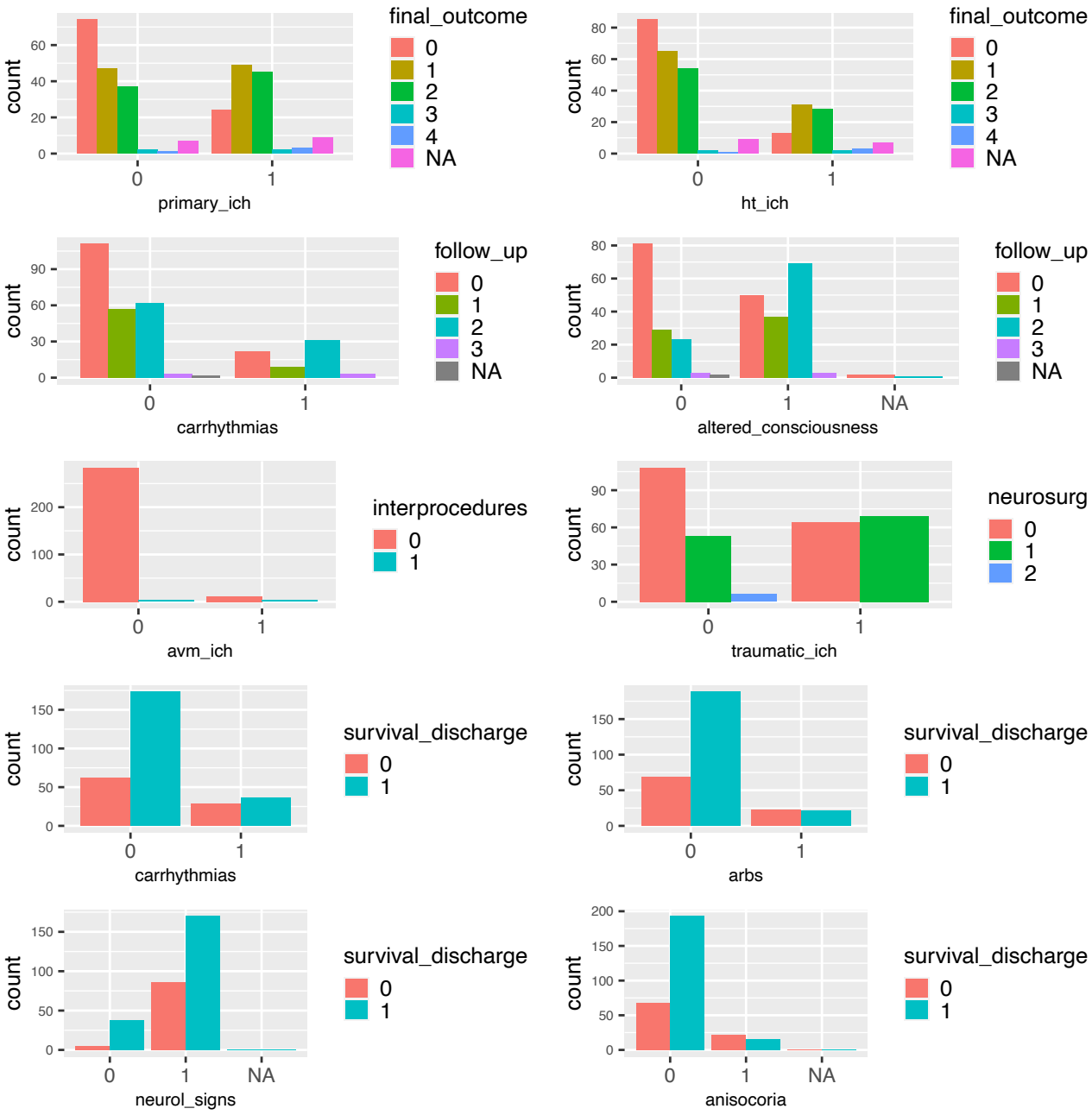


Figura D4. Diagramas de barras representando algunas de las asociaciones estadísticamente significativas más relevantes en los test χ^2 con simulación de Monte Carlo.

8.5. Anexo E: Resultados del análisis de supervivencia.

Tabla E1. Resultados significativos y cercanos a la significación en el test de Mantel-Cox (*logrank*).

Variable	*p* ajustado*
carrhythmias	0.0356
arbs	0.0272
mh_le_trauma	0.0487
other_symptoms	0.0343
neurol_signs	0.0392
anisocoria	< 0.0001
altered_consciousness	< 0.0001
tgcs	< 0.0001
primary_ich	0.0307
traumatic_ich	0.0356
ht_ich	0.0487
structural_heart_disease	0.053
diuretics	0.1007
sulfonylureas	0.053
cumarinics	0.053
mh_trauma	0.053
rlwlimb_mimpairment	0.1244

*Los valores de *p* han sido ajustados con la técnica para test múltiples FDR [44].

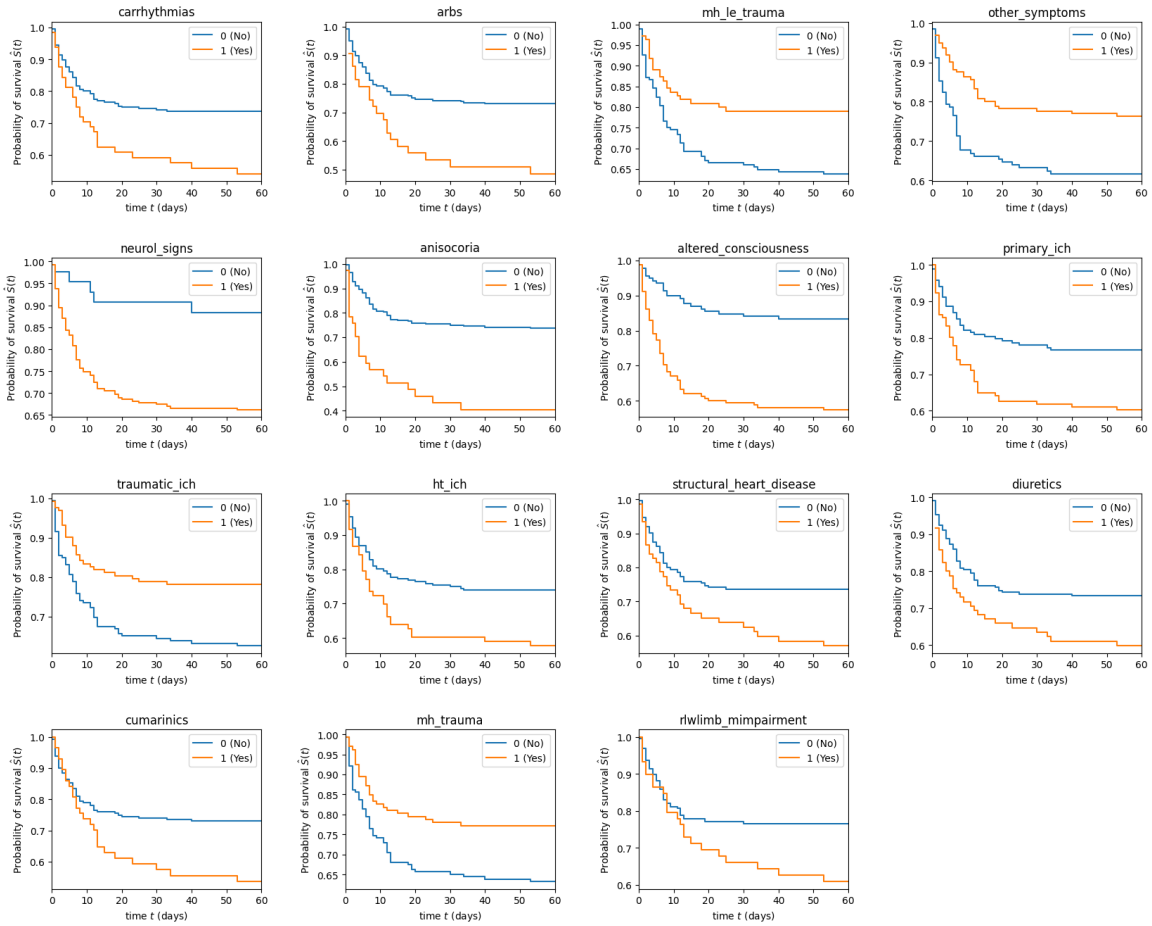


Figura E1. **Curvas de supervivencia que han resultado significativas o próximas a la significación en el test de *logrank*.**

Table E2. Variables predictoras de los modelos CPH y CoxNet

Variable
hospitalization_days*
tgcs
glucose
altered_consciousness
onset_h
prothrombin_activity

*Variable utilizada solo en los modelos CoxNet.

Table E3. Selección de variables predictores para los modelos de supervivencia SVM, SSVM, RSF, GB, AFT

Variable
sex
nfamily_medhist
ht
dmellitus
dyslipidemia
structural_heart_disease
neurological_diseases
antihypertensives
antidiabetics
hypolipidemics
anticoagulants
antiplatelets
aspirin
cumarinics
headache
emesis
visual_disturbances
seizures
neurol_signs
anisocoria
altered_consciousness
ruplimb_mimpairment
luplimb_mimpairment
rlwlimb_mimpairment
llwlimb_mimpairment
balance_impairment
tgcs
onset_h
primary_ich
glucose
urea

creatinine

potasium

prothrombin_activity

hematocrit

maxfibrinogen

age

La selección se realizó en base a los resultados estadísticos y a criterios clínicos.

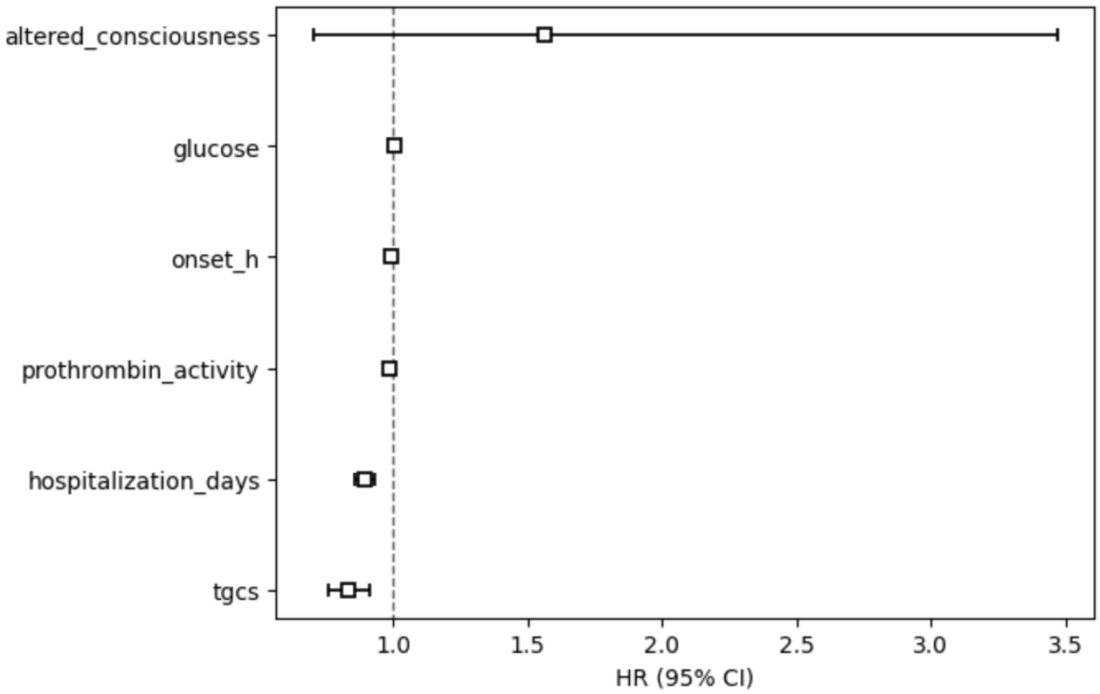


Figura E2. Valores de las hazard ratio [*exp(β)*] del modelo CPH 1.

Tabla E4. Importancia de las variables para el RSF.

	Importancia	
Variable	Media	DE
tgcs	0.072305	0.040900
onset_h	0.066667	0.040532
urea	0.032504	0.015507
glucose	0.021891	0.039543
antihypertensives	0.014925	0.004806
neurological_diseases	0.012272	0.025032
altered_consciousness	0.007297	0.007005
antiplatelets	0.006965	0.002437
antidiabetics	0.006633	0.006466
ht	0.006302	0.003382
neurol_signs	0.005307	0.003382
primary_ich	0.004312	0.004008
dmellitus	0.002985	0.003541
ruplimb_mimpairment	0.002322	0.003076
prothrombin_activity	0.002322	0.004760
aspirin	0.001990	0.003040
balance_impairment	0.001658	0.002345
emesis	0.001658	0.002345
anisocoria	0.001327	0.004248
structural_heart_disease	0.001327	0.002853
headache	0.000995	0.001990
cumarinics	0.000995	0.001990
luplimb_mimpairment	0.000332	0.002853
rlwlimb_mimpairment	0.000332	0.004248
hematocrit	-0.000332	0.017563
llwlimb_mimpairment	-0.001327	0.007144
creatinine	-0.002985	0.010562
sex	-0.003317	0.009555
age	-0.003648	0.015677
nfamily_medhist	-0.006302	0.011698
anticoagulants	-0.006633	0.004691
maxfibrinogen	-0.007297	0.019723
potasium	-0.012935	0.031924

Tabla E5. Importancia de las variables para GB *tree*

Variable	Importancia
urea	0.190349
tgcs	0.189099
onset_h	0.101046
glucose	0.093115
maxfibrinogen	0.069387
prothrombin_activity	0.056030
neurological_diseases	0.044980
potasium	0.044247
creatinine	0.034233
headache	0.031924
nfamily_medhist	0.028153
age	0.024682
hematocrit	0.015628
sex	0.014416
hypolipidemics	0.014368
antidiabetics	0.013417
dmellitus	0.008877
emesis	0.008632
llwlimb_mimpairment	0.005019
altered_consciousness	0.004894
luplimb_mimpairment	0.002504
antihypertensives	0.002016
ruplimb_mimpairment	0.001380
primary_ich	0.000867
anisocoria	0.000736
cumarinics	0.000000
ht	0.000000
dyslipidemia	0.000000
structural_heart_disease	0.000000
balance_impairment	0.000000
anticoagulants	0.000000
antiplatelets	0.000000
rlwlimb_mimpairment	0.000000
aspirin	0.000000
seizures	0.000000
visual_disturbances	0.000000
neurol_signs	0.000000

Tabla E6. Importancia de las variables para el modelo GB *component-wise least squares*

Variable	Importancia
neurological_diseases	1.36
llwlimb_mimpairment	1.11
altered_consciousness	0.62
primary_ich	0.44
sex	0.31
ruplimb_mimpairment	0.31
antidiabetics	0.18
anisocoria	0.11
structural_heart_disease	0.02
age	0.004
glucose	0.002
maxfibrinogen	- 0.001
onset_h	-0.001
emesis	-0.02
hypolipidemics	-0.03
nfamily_medhist	-0.13
tgcs	-0.19
headache	-0.37
aspirin	-0.43
visual_disturbances	-1.24

Tabla E7. Importancia de las variables para GB AFT

Variable	Importancia
tgcs	0.281055
urea	0.112091
potasium	0.078802
maxfibrinogen	0.070504
glucose	0.067010
headache	0.051992
onset_h	0.048622
prothrombin_activity	0.047155
creatinine	0.042779
llwlimb_mimpairment	0.040646
hematocrit	0.035126
age	0.027966
neurological_diseases	0.024549
nfamily_medhist	0.021362
emesis	0.012678
primary_ich	0.010446
rlwlimb_mimpairment	0.004346
sex	0.004293
hypolipidemics	0.004124
balance_impairment	0.003918
structural_heart_disease	0.002704
anisocoria	0.002179
antihypertensives	0.001346
anticoagulants	0.001229
dmellitus	0.000952
ruplimb_mimpairment	0.000758
dyslipidemia	0.000680
ht	0.000346
altered_consciousness	0.000167
luplimb_mimpairment	0.000148
antiplatelets	0.000027
seizures	0.000000
antidiabetics	0.000000
aspirin	0.000000
cumarinics	0.000000
visual_disturbances	0.000000
neurol_signs	0.000000

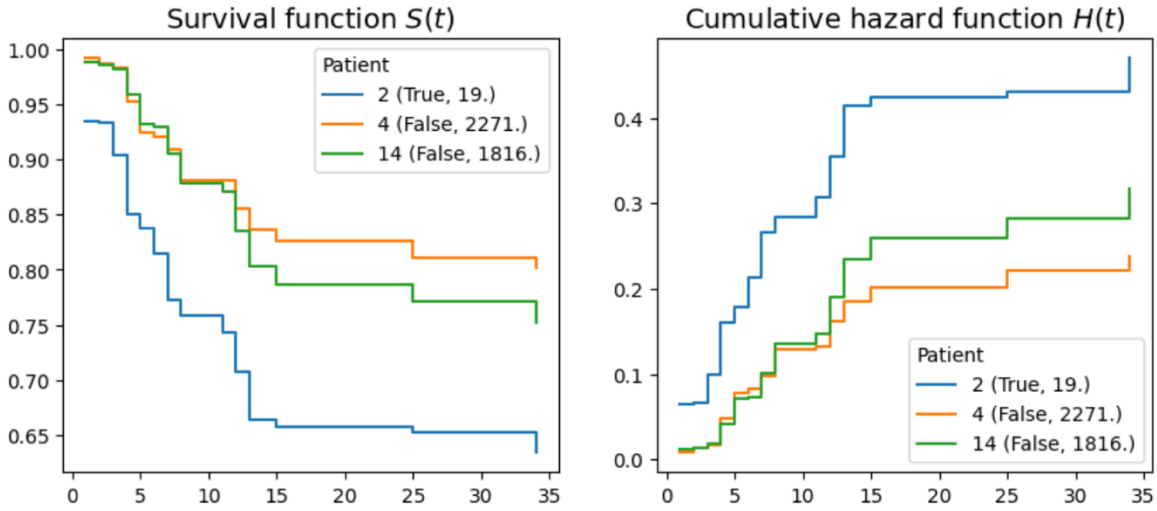


Figura E3. Ejemplo de predicciones de tres pacientes del modelo *RSF*.