

DESARROLLO Y CALIBRACIÓN DE MIMIC-CPM PARA CALIDAD DE RECUPERACIÓN POSTANESTÉSICA

SONIA M. BENÍTEZ¹, FERNANDO R. VÁZQUEZ PEÑA², RAÚL B. PUYOL³, MARÍA F. GRANDE RATTI⁴,
MARÍA VÁZQUEZ NAVARRO⁵, EULOGIO CORDÓN POZO⁶

¹Servicio de Anestesiología, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina, ²Servicio de Medicina Familiar, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina, ³Departamento de Medicina Crítica, Hospital Escuela José de San Martín, Corrientes, Argentina, ⁴Área de Investigación en Medicina Interna, Servicio de Clínica Médica, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina, ⁵Facultad de Ingeniería (FIUBA), Universidad de Buenos Aires, Buenos Aires, Argentina, ⁶Departamento de Organización de Empresas, Universidad de Granada, España

Dirección postal: Sonia M. Benítez, Servicio de Anestesiología, Hospital Italiano de Buenos Aires, Tte. Gral. Juan D. Perón 4190, 1199 Buenos Aires, Argentina

E-mail: soniamabel.benitez@hospitalitaliano.org.ar

Recibido: 7-VI-2026

Aceptado: 9-X-2026

Resumen

Introducción: Los modelos de predicción clínica (MPC) permiten personalizar la atención perioperatoria, pero su desarrollo desde ecuaciones estructurales con variables latentes es infrecuente. Nuestro objetivo fue construir y calibrar un modelo para predecir individualmente la calidad de recuperación postanestésica (CRPA), partiendo de un modelo estadístico MIMIC de 13 ítems validado previamente.

Materiales y métodos: Utilizando datos de 318 pacientes (del estudio de validación del QoR-15 al español rioplatense), implementamos el marco MIMIC-CPM para transformar los coeficientes estructurales en predicciones individuales calibradas. El proceso incluyó: (a) promedio ponderado por cargas factoriales de segundo orden, (b) calibración ajustando pendiente (Factor K) e intercepto basal (b0), (c) validación por subgrupos con intervalos de confianza del 95%, y (d) conversión a probabilidades por cuantiles. Comparamos tres metodologías de estimación: SEM estructural, SEM empírico (Steyerberg) y SAM Local.

Resultados: Se conformaron diez perfiles clínicos ($n \geq 15$). Todos los valores observados cayeron dentro del margen de error (IC 95%) de sus respectivas predicciones. La convergencia entre la variabilidad teórica del

modelo (1.304) y la empírica (1.613) arrojó una diferencia moderada y esperable ($\Delta = 0.309$), confirmando un ajuste estadístico robusto. Los tres motores estadísticos produjeron resultados congruentes. La validación interna (validación cruzada de 10 pliegues y bootstrapping $B = 500$) no evidenció sobreajuste (optimismo $R^2 = -0.001$; $K = -0.007$ y $b0 = 0.035$).

Conclusión: El marco metodológico MIMIC-CPM constituye una aproximación estadística robusta y reproducible a nivel local para la predicción de la calidad de recuperación postanestésica, integrando la estructura factorial de 13 ítems del cuestionario QoR-15 de forma directa.

Palabras clave: periodo de recuperación de la anestesia, modelos de predicción clínica, QoR-15, modelos estadísticos, calibración, análisis de sensibilidad

Abstract

Development and calibration of MIMIC-CPM for post-anesthetic recovery quality

Introduction: Clinical prediction models (CPMs) enable personalized perioperative care, yet their devel-

opment from structural equation models with latent variables is uncommon. We aimed to build and calibrate an individual predictive framework for quality of post-anesthetic recovery (QoR) from a previously validated 13-item MIMIC model.

Materials and methods: Using 318 patients from the QoR-15 Rioplatense Spanish validation study, we implemented the MIMIC-CPM framework to transform structural coefficients into calibrated individual predictions. The procedure comprised: (a) weighted average calculation using second-order factor loadings, (b) calibration via K-Factor and Intercept b_0 , (c) subgroup validation with 95% confidence intervals, and (d) probabilistic conversion into Gaussian quartiles. Three estimation engines were compared: structural SEM, empirical SEM (Steyerberg), and Local SAM.

Results: Ten clinical profiles ($n \geq 15$) were formed. All observed values fell within the 95% CIs of their calibrated predictions. Structural-empirical convergence yielded a moderate difference ($\delta = 0.309$) between the theoretical (1.304) and empirical (1.613) standard errors. All three engines produced congruent predictions in the sensitivity analysis. Internal validation (10 plieues cross-validation and bootstrapping $B = 500$) showed no evidence of overfitting (optimism in $R^2 = -0.001$; in $K = -0.007$ and $b_0 = 0.035$).

Conclusion: The MIMIC-CPM framework is a robust and reproducible approach at the local level for individual QoR prediction, providing an operational clinical model derived directly from the validated 13-item QoR-15 factorial structure.

Key words: anesthesia recovery period, clinical prediction rules, QoR-15, statistical models, calibration, sensitivity analysis

PUNTOS CLAVE Conocimiento actual

- Los modelos de predicción clínica tradicionales en anestesiología suelen basarse en regresiones que emplean la sumatoria simple de puntajes brutos.
- Este enfoque convencional asume de forma incorrecta que todos los ítems de un cuestionario poseen el mismo peso matemático e ignora el sesgo inherente del error de medición individual.

Contribución del artículo al conocimiento actual

- Introduce el marco MIMIC-CPM, transformando un modelo de variables latentes en una herramienta de predicción individual calibrada, con convergencia estructural-empírica evaluada mediante la diferencia entre error teórico y empírico ($\delta = 0.309$).
- Se ofrece un *dashboard* clínico interactivo de acceso abierto para estimar la calidad de recuperación pos anestésica en el punto de atención.

Los modelos de predicción clínica (MPC) son herramientas cuantitativas que permiten estimar qué probabilidad tiene un paciente en particular de experimentar un desenlace específico, combinando múltiples variables predictoras en una ecuación calibrada¹. En el ámbito de la cirugía y la anestesia, poder anticipar cómo será la calidad de recuperación postanestésica (CRPA) de un paciente permitiría optimizar su plan anestésico, identificar a quienes tienen un riesgo elevado de una mala recuperación y tomar medidas preventivas. Esto es especialmente valioso sabiendo que actualmente existen cuestionarios validados para medir esta recuperación de manera estandarizada y validada².

Sin embargo, la mayoría de los modelos predictivos en anestesiología se construyen utilizando regresiones simples o logísticas basadas en la suma cruda de puntajes. El problema de este enfoque es que asume implícitamente que todas las preguntas de un cuestionario pesan lo mismo y están libres de errores de medición, lo que genera imprecisión³. Para solucionar esto, existen métodos más avanzados como los modelos de ecuaciones estructurales (SEM), y particularmente los modelos MIMIC (*Multiple Indicators, Multiple Causes*), los cuales operan sobre "variables latentes". Esto significa que son capaces de depurar el error de medición y reflejar el verdadero peso de cada variable^{3,4}.

En un artículo complementario enviado a esta misma revista, validamos la versión en español rioplatense del cuestionario QoR-15⁵. La versión

definitiva quedó conformada por 13 ítems agrupados en tres dimensiones (Bienestar, Competencia oral y Dolor), modelados a través de un esquema MIMIC jerárquico que incluyó varios predictores clínicos significativos⁵ (Anexo 1). El objetivo del presente trabajo es dar el siguiente paso: tomar ese modelo estructural complejo y transformarlo en una herramienta clínica práctica que permita predecir la recuperación a nivel individual, a la que denominamos MIMIC-CPM (Clinical Prediction Model), aplicando técnicas de calibración y validación recomendadas internacionalmente^{1,6}.

Materiales y métodos

Diseño, muestra y aspectos éticos

Este estudio analítico fue autorizado por el Hospital Italiano de San Justo y el Hospital Escuela Gral. San Martín de Corrientes (Servicios de Anestesiología, Directores y Comité de Ética de Protocolos de Investigación, protocolo N° 7273). Se obtuvo el consentimiento informado de los participantes para el uso anónimo y confidencial de sus datos con fines de investigación, cumpliendo con la Ley de Habeas Data N° 25.326 y la Declaración de Helsinki.

Se utilizaron los datos de los 318 pacientes del estudio de validación del QoR-15⁵, cuyas características clínicas y demográficas han sido reportadas previamente. Para la construcción del modelo predictivo, se seleccionaron las cuatro variables que demostraron significación estadística en el modelo MIMIC: clasificación ASA (*American Society of Anesthesiologists*), edad en *splines*, duración de la cirugía en minutos y puntaje prequirúrgico. Es importante destacar que, si bien en el artículo de validación de la escala se utilizó el puntaje prequirúrgico original de 15 ítems como criterio externo, para la construcción de este modelo predictivo clínico (MIMIC-CPM) se utilizó el puntaje basal depurado de 13 ítems. Esto procura una aplicación clínica coherente, circunscribiendo el modelo predictivo a los 13 ítems de nuestra validación local⁵. El paciente completa esta versión optimizada del cuestionario antes de la cirugía, y el sistema utilizará dicha línea de base para predecir su recuperación. Con el objetivo de optimizar la especificación estructural y capturar la relación de cada variable con la calidad de recuperación, se compararon cuatro modelos alternativos, variando la forma funcional de la edad y la duración quirúrgica. Teniendo en cuenta fundamentalmente los valores de AIC (*Akaike Information Criterion*), el porcentaje de varianza explicada y la parsimonia, se seleccionó como mejor configuración la del

modelo que incorpora la edad mediante *splines* cúbicos ($df=3$) y la duración quirúrgica en minutos lineal⁵.

Marco MIMIC-CPM: arquitectura general

El marco predictivo se construyó en las siguientes fases secuenciales. Para evaluar la robustez de los resultados (análisis de sensibilidad), cada fase se ejecutó en paralelo utilizando tres motores de estimación: (1) SEM estructural con estimador robusto de máxima verosimilitud (MLR), que proporciona el error estándar teórico del modelo; (2) SEM empírico (Steyerberg), que adopta los mismos coeficientes estructurales pero reemplaza el error teórico por el error estándar empírico de los residuos, siguiendo la metodología de validación de modelos predictivos⁶; y (3) SAM Local (Structural After Measurement)⁷, que estima secuencialmente la parte de medición y la parte estructural del modelo, proporcionando una verificación independiente de la arquitectura factorial.

Fase 1: Promedio ponderado

En lugar de estimar los puntajes factoriales (*factor scores*), los cuales están sujetos a indeterminación factorial y dificultan la interpretación directa, implementamos el cálculo de un “promedio ponderado” (o puntuaciones compuestas observadas) basado en las cargas factoriales de un modelo de segundo orden^{8,9}. Concretamente, este método calcula, para cada paciente, la suma de los productos entre el valor de cada uno de los 13 ítems (Anexo 2) y el producto de sus cargas factoriales (nivel 1 x nivel 2) factoriales (nivel 1 x nivel 2), normalizando el resultado a la escala original (0 a 10) mediante la división por la suma de dichos productos.

Cada motor de estimación utilizó sus propias cargas factoriales extraídas de manera independiente, asegurando que las tres vías funcionaran de forma autónoma.

Fase 2: Calibración a nivel individual

Para reducir el riesgo de sesgo de agregación y evitar el sesgo de agregación (falacia ecológica), la calibración se realizó estrictamente a nivel individual ($N = 318$), siguiendo las recomendaciones metodológicas actuales para modelos de predicción clínica^{1,6}. A partir de los valores observados (ver Fase 1) y las predicciones crudas estructurales (sumatoria del producto de coeficientes por el valor de los predictores), se calculó: (a) el Factor K (*Calibration Slope*) mediante la pendiente de una regresión de mínimos cuadrados ordinarios (OLS), y (b) el Intercepción b_0 (*Calibration-in-the-large*) para corregir el desajuste sistemático. La predicción individual calibrada se obtuvo aplicando la ecuación: Predicho_calibrado = (Predicho_crudo x Factor K) + Intercepción b_0 .

Fase 3: Validación gráfica por subgrupos e intervalos de confianza

Con el propósito de evaluar empíricamente la precisión de las predicciones calibradas en distintos escenarios, se conformaron subgrupos representativos garantizando un tamaño muestral mínimo ($n \geq 15$) para la estabilidad de las estimaciones^{10,11}. Para ello, se dicotomizaron transitoriamente las variables: la duración de la cirugía y el puntaje prequirúrgico según sus medianas (debido a su asimetría), y la edad por su media, combinándose con la clasificación ASA. Este procedimiento permitió estructurar diez perfiles clínicos heterogéneos (Fig. 1). Cabe destacar que esta segmentación se aplicó exclusivamente para la validación visual; en el modelo predictivo, todas las variables numéricas operan de forma continua. Para cada perfil, se comparó la media de las puntuaciones compuestas observadas (“rombo rojo”) con la media de las predicciones calibradas (“círculo celeste”). Se consideró concordancia adecuada cuando la media observada de los pacientes caía dentro del intervalo de confianza (banda gris) construido a partir del error estándar específico de cada motor de estimación evaluado: SEM Estructural, SAM Local y SEM empírico de Steyerberg⁶. Para mantener una confianza global del 95% a través de todas las evaluaciones simultáneas (10 perfiles clínicos evaluados de forma cruzada bajo las dos arquitecturas de modelado principales: SEM y SAM, conformando 20 comparaciones) y evitar la inflación del error tipo I, se aplicó una corrección conservadora tipo Bonferroni ($Z = 3.023$).

Fase 4: Conversión probabilística

Cada predicción calibrada se interpreta como el centro de una distribución gaussiana cuya dispersión depende del margen de error del modelo. Utilizando las 318 predicciones calibradas de nuestra muestra, definimos umbrales para la conformación de los cuatro cuartiles (Q_1 , Q_2 y Q_3 ; percentiles 25, 50 y 75, respectivamente). Mediante funciones de probabilidad acumulada (*pnorm*), calculamos la probabilidad que tiene un paciente específico de pertenecer a cada cuartil (desde una recuperación pobre hasta una excelente).

Fase 5: Validación estructural (convergencia de residuos)

Siguiendo las recomendaciones de Steyerberg⁶, calculamos el error estándar “empírico” o la desviación estándar de los residuos empíricos (definida como la desviación estándar de la diferencia individual entre el valor real observado y la predicción calibrada para los 318 casos) y la comparamos con el error estándar teórico

calibrado (corregido por el factor K) de nuestro modelo estructural. Si ambos valores son muy similares (un delta o diferencia absoluta mínima), significa que nuestro modelo predice de forma sumamente realista.

Fase 6: Predicciones individuales y dominio de aplicabilidad

Si bien la validación gráfica se realizó agrupando a los pacientes en perfiles específicos, a nivel clínico el modelo arroja predicciones de carácter estrictamente individual. A la ecuación matemática se le introducen los valores numéricos exactos del paciente (ej., 75 minutos de duración, 48 años, puntaje previo de 125 y ASA II). Estos datos alimentan una distribución gaussiana específica cuya dispersión varía según el error estándar propio de cada motor de estimación (error estructural calibrado para SEM, error estándar empírico para SEM-Steyerberg, y error estructural calibrado para SAM Local). Al aplicar los umbrales correspondientes a cada modelo, se generan cuartiles similares, pero estadísticamente independientes, culminando en un pronóstico probabilístico totalmente personalizado (ver *dashboard* clínico interactivo).

Dominio de aplicabilidad y límites de seguridad

Para limitar el riesgo de extrapolación del *dashboard* interactivo y prevenir el sesgo por extrapolación (un riesgo inherente al proyectar variables más allá de la muestra de entrenamiento), se establecieron restricciones de frontera para las variables de entrada, delimitadas estrictamente por la densidad empírica de la cohorte de derivación. En primer lugar, la duración quirúrgica se truncó en un máximo de 240 minutos (4 horas); si bien se observaron casos atípicos de hasta 8 horas, el 99% de las intervenciones finalizaron antes de los 235 minutos, por lo que este tope evita que los outliers arrastren linealmente la predicción. El puntaje prequirúrgico basal se delimitó entre 60 y 130 puntos, respetando el techo matemático absoluto del instrumento y un límite inferior de 60 puntos (en coherencia con el valor mínimo absoluto de 57 registrado empíricamente en la cohorte). Adicionalmente, para evitar que los *splines* cúbicos generen predicciones inestables en colas poblacionales de baja densidad, se estableció un piso global de 17 años (coincidente con el paciente más joven de la muestra) y una edad máxima admitida que se definió por subgrupo aproximándose al percentil 95 (P95) de cada subgrupo: se establecieron topes de 55 años para ASA I (P95 real = 53.0), 70 años para ASA II (P95 real = 70.2) y 85 años para ASA III (P95 real = 83.3). Finalmente, aunque el modelo estructural se estimó con la cohorte completa ($n=318$), el ingreso en la interfaz

clínica se restringió a las categorías ASA I, II y III, excluyendo precautoriamente a la categoría ASA IV (N=3) debido a que carece de la densidad muestral necesaria para ofrecer estimaciones individuales estables en perfiles de alto riesgo.

Fase 7: Análisis de sensibilidad

Los tres motores de estimación se ejecutaron en paralelo en un *dashboard* interactivo. Esto nos permitió verificar que, ante un mismo caso clínico, las predicciones individuales y las probabilidades por cuartil fuesen similares sin importar el método estadístico de fondo.

Prevención de sobreajuste y estrategia de validación interna

Una limitación frecuentemente señalada en los modelos de predicción clínica desarrollados sobre muestras únicas es el riesgo de sobreajuste (*overfitting*), por el cual los parámetros se optimizan sobre los mismos datos utilizados para evaluarlos, inflando artificialmente los índices de bondad de ajuste^{1,12}. Esta limitación es especialmente relevante en el modelo propuesto, que incorpora *splines* cúbicos restringidos para la edad (df=3, nudos en 38 y 53 años), siendo más probable que se produzca un sobreajuste en modelos con mayor complejidad paramétrica.

Para abordar esta limitación de forma empírica, se implementó una estrategia de validación interna combinada siguiendo las recomendaciones de Steyerberg⁶ y las guías TRIPOD¹³. Se llevó a cabo una validación cruzada de 10 pliegues estratificada por clasificación ASA (validación cruzada de k-pliegues), complementada con *bootstrapping* de 500 re-muestras para la estimación y corrección formal del optimismo¹⁴. La estratificación por ASA garantiza la representación proporcional de los perfiles clínicos en cada pliegue, siguiendo el criterio de partición informada recomendado para muestras de tamaño moderado¹⁵.

Un aspecto crítico de la implementación en nuestro modelo es la regeneración de los términos del *spline* en cada partición con los nudos del modelo global (38 y 53 años), garantizando que la forma funcional sea idéntica en cada *fold* y en cada re-muestra *bootstrap*. Sin esta precaución, cada subconjunto generaría nudos propios, produciendo predicciones inconsistentes entre pliegues.

Modelo predictivo: programas y tecnologías

Se utilizó RStudio para los *scripts* de análisis, y tecnología HTML/JavaScript para construir el *dashboard* clínico interactivo. Este se encuentra disponible con acceso abierto en <https://radiant-monstera-865a81.netlify.app>.

Resultados

Calibración por subgrupos por subgrupos y validación por intervalos de confianza

La combinación de las variables predictoras permitió conformar diez perfiles clínicos (que abarcan desde el escenario más favorable hasta el de peor pronóstico esperado). La Figura 1 muestra esta validación comparando simultáneamente los tres motores de estimación (SEM Estructural, SEM Steyerberg y SAM Local). Como puede observarse en los tres paneles, las predicciones calibradas promedio del modelo (círculos celestes) se alinean estrechamente con los puntajes medios observados en los pacientes (rombos rojos) para todos los subgrupos. De manera consistente, los valores empíricos cayeron sistemáticamente dentro de los márgenes de error predictivo del 95% (bandas grises horizontales, ajustadas con corrección de Bonferroni para comparaciones múltiples), confirmando un ajuste estadístico robusto en todas las franjas de riesgo y demostrando que las tres metodologías producen resultados congruentes.

Convergencia estructural-empírica

La diferencia entre el error teórico calibrado (1.304) y el empírico (1.613) fue de 0.309 puntos, lo que indica que el error estructural subestima moderadamente la dispersión observada. Por este motivo se incorporó el motor SEM-Steyerberg, que utiliza el error empírico y proporciona estimaciones probabilísticas más conservadoras.

Ejemplo de predicción clínica individual

Si bien la validación gráfica se realizó agrupando a los pacientes en perfiles específicos, a nivel clínico el modelo arroja predicciones de carácter estrictamente individual. Al introducir los valores numéricos exactos de un sujeto particular, por ejemplo, un paciente de 65 años, con una duración quirúrgica de 120 minutos, clasificación ASA III y un puntaje prequirúrgico basal de 125 puntos, el sistema alimenta una distribución gaussiana personalizada. Esto permite determinar un pronóstico probabilístico específico y a medida para ese caso único (Fig. 2).

Figura 1 | Calibración por subgrupos clínicos y motores de estimación. En el eje vertical del panel izquierdo se detallan los 10 perfiles clínicos principales, contruidos cruzando edad (J=Joven, M=Mayor), estado físico (A1=ASA I, A2=ASA II, A3=ASA III), duración quirúrgica (C=Corta, L=Larga) y puntaje basal (A=Alto, B=Bajo). El tamaño muestral (n) de cada estrato se indica entre paréntesis (N total graficado = 248). Los 70 pacientes restantes de la cohorte original (N = 318) pertenecían a combinaciones clínicas infrecuentes (N < 15) que fueron excluidas de este gráfico para preservar la estabilidad estadística de los intervalos empíricos. Los perfiles se ordenan desde el peor pronóstico esperado (abajo) hasta el mejor (arriba). Los círculos celestes indican la predicción calibrada promedio del modelo y los rombos rojos la media real observada. Las bandas horizontales grises representan intervalos de confianza del 95% simultáneos (corrección de Bonferroni para 20 comparaciones cruzadas). En todos los perfiles y motores de estimación, los valores observados caen sistemáticamente dentro del margen de error teórico.

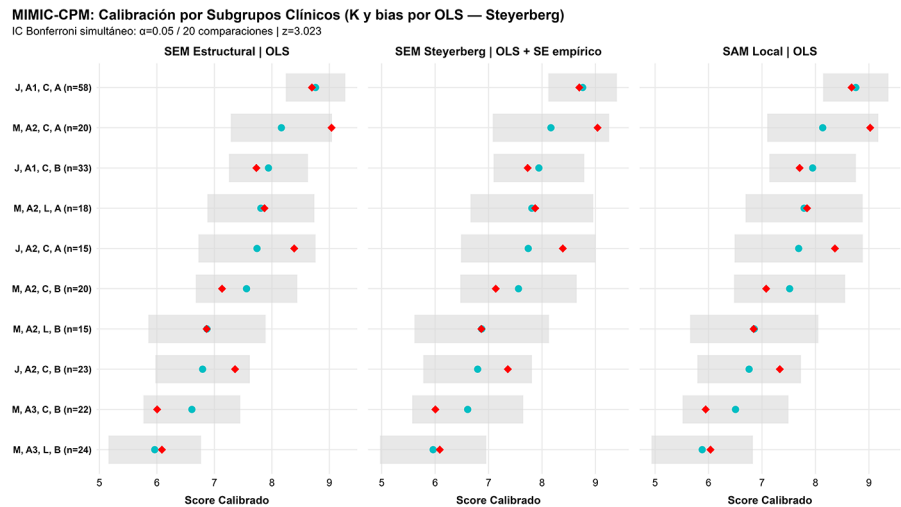
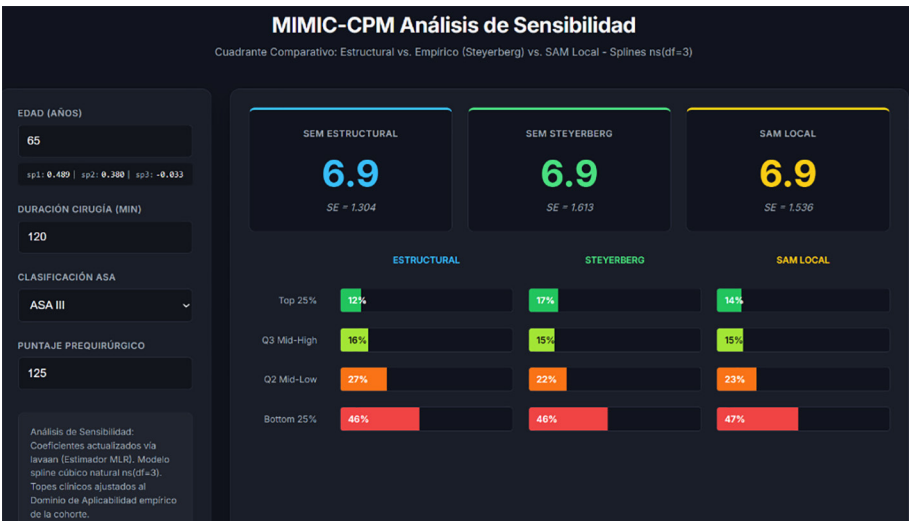


Figura 2 | Interfaz del Dashboard MIMIC-CPM: visualización del análisis de sensibilidad triple. Predicción individual calibrada y distribución de probabilidades por cuartiles para los tres motores de estimación comparados (SEM Estructural, SEM Steyerberg y SAM Local).



Análisis de sensibilidad

Al someter este mismo escenario clínico a los tres motores estadísticos evaluados, los resultados evidenciaron una notable equivalencia y

robustez entre modelos. Para el caso de referencia citado, la métrica de puntuación ponderada (score) calibrada fue idéntica en todas las plataformas: 6.9 (SEM Estructural, SEM Steyerberg y

SAM Local). En términos de riesgo probabilístico, tanto el enfoque estructural como el empírico (SEM) estimaron un 46.0% de probabilidad de que el paciente pertenezca al cuartil de peor recuperación (cuartil inferior), mientras que el modelo SAM Local computó un 47.0%. Estas discrepancias marginales entre metodologías independientes sugieren estabilidad y consistencia interna del marco predictivo desarrollado.

Ajuste del modelo y calibración “aparente”

El modelo convergió sin advertencias con los siguientes índices de ajuste: CFI robusto=0.902 (ajuste aceptable, contextualizado por la complejidad de un modelo de segundo orden y la penalización del estimador MLR), RMSEA robusto=0.074, SRMR=0.053, AIC=17029.2 y R^2 predictivo estructural latente de 0.455. La alineación de escalas entre el factor score estandarizado y la escala 0-10 del promedio ponderado alcanzó una fuerte concordancia con un R^2 de alineación de 0.921, un intercepto=0.967 y una pendiente=1.358. Los parámetros (que actúan como factor de traslación de escala y estimadores de calibración aparente, surgidos de comparar las predicciones individuales calibradas con los promedios ponderados individuales) fueron: $K=1.358$, Intercepto $b_0=0.967$, RMSE=0.563, R^2 de alineación=0.921 y SE=0.565 (Tabla 1). La reducción controlada del R^2 de alineación respecto al modelo lineal ($\Delta R^2=0.015$) representa el precio justificado de corregir la no-linealidad intrínseca de la edad y la distorsión secundaria sobre la duración quirúrgica. La Tabla 2 cuantifica este *trade-off*.

Nuestros resultados muestran una estabilidad interna adecuada del modelo propuesto tras la corrección del supuesto de linealidad.

En la validación cruzada de 10 pliegues (Tabla 3), el Factor K mostró una variabilidad mínima entre pliegues (media=1.357, SD=0.040, rango: 1.280 - 1.430), evidenciando que la pendiente de calibración es con baja variabilidad ante diferentes particiones de la muestra¹⁶. El intercepto b_0 mostró mayor sensibilidad a la composición de los pliegues (SD = 0.732), estabilizándose en la estimación global por *bootstrapping* (b_0 corregido = 0.932). El R^2 de alineación OOS fue de 0.915 (SD=0.040) (Tabla 3). El *bootstrapping* con $B=500$ re-muestras cuantificó formalmente el optimismo de nuestro modelo: -0.001 en R^2 (fluctuación estocástica propia del re-muestreo que evidencia un optimismo nulo), -0.007 en Factor K y 0.035 en Intercepto b_0 (Tabla 4).

La ausencia de optimismo en los parámetros de calibración es particularmente relevante, ya que implica que la pendiente y el nivel de calibración son igualmente válidos en datos no vistos^{6,12}. Las métricas corregidas por optimismo (*bootstrap*) fueron R^2 de alineación=0.922, RMSE=0.566 y $K=1.365$ (Tabla 4), valores que demuestran una alta consistencia con los aparentes. La mayor variabilidad entre *folds* respecto al modelo con edad lineal (SD $R^2=0.040$ vs 0.022) refleja la complejidad adicional del *spline* y es esperable. El *fold* 8 ($R^2=0.834$, $n=31$) representó el caso menos favorable, atribuible a la distribución de perfiles etarios en esa partición específica, no a una inestabilidad estructural del modelo. En conjunto, la convergencia entre las tres estimaciones, aparente ($R^2=0.921$), OOS ($R^2=0.915$) y corregida por *bootstrap* ($R^2=0.922$), refuta empíricamente la hipótesis de sobreajuste y respalda la solidez del modelo propuesto como herramienta de predicción individual calibrada.

Tabla 1 | Comparativa aparente, OOS y *Bootstrap* corregido

Método	K	Intercepto b_0	RMSE	R^2 de alineación	SE
Aparente (n=318)	1.358	0.967	0.563	0.921	0.565
validación cruzada de 10 pliegues OOS	1.357	0.975	0.561	0.915	0.580
Bootstrap corregido (B=500)	1.365	0.932	0.566	0.922	0.568
Optimismo (bootstrap medio)	-0.007	0.035	-0.003	-0.001	-0.003

K = calibration slope (Factor *K*). Intercepto b_0 = calibration-in-the-large. RMSE = raíz del error cuadrático medio. R^2 = coeficiente de determinación. SE = error estándar de los residuos. OOS = out-of-sample. Optimismo = métrica aparente en bootstrap menos métrica del bootstrap aplicado a datos originales. Negrita = optimismo nulo o clínicamente irrelevante. Cursiva = optimismo presente pero aceptable (RMSE: -0.003).

Tabla 2 | Comparación modelo con edad continua vs. modelo con *spline*

Métrica	Modelo 1	Modelo 2	Diferencia	Observaciones
R ² alineación aparente	0.936	0.921	-0.015	Coherente al corregir supuesto linealidad
R ² de alineación OOS (10 pliegues)	0.932	0.915	-0.017	Esperado con <i>spline</i>
R ² alineación corregido (boot)	0.934	0.922	-0.012	Coherente y aceptable
Optimismo R ² (alineación)	0.002	-0.001	-0.003	Sin sobreajuste
Optimismo K	0.000	-0.007	-0.007	Calibración estable
RMSE aparente	0.507	0.563	+0.056	Coherente con aumento complejidad
Supuestos violados	2	0	-2	Modelo 2 metodológicamente correcto

La ligera reducción del R² de alineación de escalas (0.921 vs. 0.936, ver Tabla 2) es un artefacto matemático previsible en la estimación de factor scores al incorporar términos no lineales. Sin embargo, en términos de verdadera capacidad predictiva, la incorporación del *spline* mejoró sustancialmente el ajuste global (ANOVA lineal vs. *spline*: $p < 0.001$, ΔAIC *spline* = -5.6) y aumentó la varianza explicada del constructo latente (R² predictivo estructural de 0.423 a 0.455). Esta mejora en el ajuste, combinada con la necesidad de corregir los dos supuestos violados, justifica estadísticamente que el Modelo 2 sea considerado el metodológicamente correcto.

Tabla 3 | Detalle por fold (10 pliegues cross-validation)

Fold	N test	K	Intercepto b0	RMSE OOS	R ² de alineación OOS	SE OOS
1	32	1.364	2.245	0.685	0.894	0.671
2	32	1.356	0.885	0.529	0.952	0.533
3	31	1.381	-0.197	0.622	0.912	0.632
4	32	1.332	1.270	0.405	0.964	0.392
5	32	1.363	0.107	0.685	0.892	0.696
6	32	1.373	0.941	0.478	0.939	0.483
7	32	1.376	0.996	0.589	0.884	0.595
8*	31	1.318	1.484	0.603	0.834	0.613
9	32	1.430	0.410	0.547	0.932	0.532
10	32	1.280	1.613	0.469	0.951	0.476
Media	---	1.357	0.975	0.561	0.915	0.580
SD	---	0.040	0.732	0.093	0.040	0.096

Estratificación por clasificación ASA. K e Intercepto b0 calculados en el conjunto de entrenamiento (~286 casos). RMSE OOS, R² de alineación OOS y SE OOS evaluados en el pliegue de prueba (out-of-sample). * = R² de alineación OOS < 0.88 (fold 8, n=31). El fold 8 produjo el R² mínimo (0.834), atribuible a la distribución desfavorable de perfiles etarios en esa partición de 31 casos. No indica inestabilidad estructural del modelo.

Tabla 4 | Optimismo estimado por *bootstrapping* (B=500)

Parámetro	Aparente	Corregido	Optimismo medio
K (calibration slope)	1.358	1.365	-0.007
Intercepto b0	0.967	0.932	0.035
RMSE	0.563	0.566	-0.003
R ² de alineación	0.921	0.922	-0.001
SE	0.565	0.568	-0.003

Optimismo = métrica aparente en bootstrap menos métrica del bootstrap aplicado a datos originales (método Steyerberg/Harrell). Optimismo en K = -0.007 y b0 = 0.035: los parámetros de calibración son perfectamente estables. Optimismo en R² = -0.001 (fluctuación estocástica propia del re-muestreo, denotando optimismo nulo). * = optimismo nulo o mínimo. ** = optimismo presente pero aceptable (RMSE: -0.003 puntos en escala 0–10).

Con respecto al tamaño de efecto real, y a la estabilidad del modelo en ese sentido, se realizaron las mismas pruebas, validación cruzada de 10 pliegues OOS y Bootstrap corregido (B=500), ya comentadas en este apartado, comparando, mediante el cuadrado del coeficiente de Pearson, la predicción calibrada (corregida por el Factor K y el Intercepto b0) individual, y el promedio ponderado de cada uno de los 318 casos. El modelo “aparente” tuvo un R^2 de 0.355; validación cruzada de 10 pliegues (Fuera de Muestra) 0.342, y el Bootstrap Corregido (B=500) 0.331. Esto equivale a un tamaño de efecto de $f^2 = 0.550$ (calculado como $0.355 / [1 - 0.355]$), superando los tres valores el umbral de $f^2 \geq 0.35$ para ser clasificado como un efecto grande según las guías de Cohen¹⁷. Asimismo, el tamaño de efecto con el R^2 predictivo corregido por *bootstrap* fue de $f^2 = 0.495$, reafirmando la magnitud estadística del efecto.

Discusión

Este trabajo muestra que es factible que es posible tomar un modelo matemático abstracto (un modelo estructural con variables latentes) y convertirlo en una herramienta predictiva de uso clínico diario. Este “puente” metodológico es infrecuente en la literatura perioperatoria, pero ofrece una potencial ventaja: nuestras predicciones se basan en constructos cuyos coeficientes estructurales se estiman controlando el error de medición de los ítems del error de medición de los ítems. Esto presenta ventajas metodológicas frente a los modelos de regresión convencionales, que presentan la limitación inherente de simplemente sumar puntajes brutos³.

El desarrollo de nuestro marco MIMIC-CPM se fundamenta en tres pilares metodológicos. El primero es el uso del promedio ponderado por cargas factoriales^{3,9}, el cual asegura que los ítems más representativos del cuestionario tengan un impacto proporcionalmente mayor en el puntaje, algo que se pierde al hacer simples sumatorias. El segundo pilar es la técnica de calibración de Steyerberg^{1,6}, que utiliza el Factor K (para ajustar la inclinación de los puntajes) y el Intercepto b0 (para corregir el desajuste sistemático). Podemos pensar en esta calibración como si quisiéramos convertir precios de una moneda extranjera a la nuestra: el Factor K se-

ría la ‘tasa de cambio’ y el Bias representaría un ‘carga o arancel fijo’. Siguiendo con esta comparación intuitiva, el error estándar del modelo también debe ser calibrado por el Factor K; de lo contrario, intentar comparar el error predictivo sería tan engañoso como evaluar diferencias de precios utilizando monedas distintas sin haber hecho la conversión. Sin esta traducción obligatoria, los números matemáticos del modelo carecerían por completo de sentido clínico. Finalmente, el tercer pilar es la validación mediante intervalos de confianza⁶, que sugiere una calibración adecuada en una variedad de perfiles clínicos diferentes.

Nuestra decisión de agrupar temporalmente a los pacientes (categorizando las variables continuas), mientras las predicciones individuales se modelan de manera continua, representó un equilibrio razonable entre simplificar los datos para probar la calibración y conservar el detalle clínico individual. Establecer grupos de calibración con al menos 15 pacientes mostró empíricamente ser apropiado, y permitió conformar diez perfiles clínicos.

La diferencia entre el error teórico calibrado (1.304) y el empírico (1.613) fue de 0.309 puntos, lo que indica que el error estructural subestima moderadamente la dispersión observada. Por este motivo se incorporó el motor SEM-Steyerberg, que utiliza el error empírico y proporciona estimaciones probabilísticas más conservadoras.

El hecho de que tres metodologías estadísticas de estimación diferentes arrojen resultados equivalentes aporta un “seguro de calidad” adicional. Garantiza al médico tratante que el pronóstico que está viendo no es simplemente un “artefacto” o espejismo generado por usar un software específico, sino que es una propiedad robusta de los datos.

En la práctica clínica, la principal utilidad de este desarrollo se materializa a través del *dashboard* interactivo. Frente a modelos predictivos existentes en anestesiología —que a menudo se basan en desenlaces dicotómicos (complicación sí/no) o en escalas de riesgo general preoperatorio (ej. calculadoras de riesgo cardiovascular)—, este modelo podría aportar un valor incremental al enfocarse específicamente en la calidad de la recuperación post-anestésica percibida por el

propio paciente. El anestesiólogo ingresa los datos de su paciente y obtiene instantáneamente: (a) un puntaje predicho calibrado, y (b) su probabilidad de cursar una recuperación de pobre a excelente.

Con este panorama anticipado en mano, se abren escenarios clínicos concretos para modificar la conducta del profesional. Por ejemplo, ante una predicción de mala recuperación, el anestesiólogo podría optar por técnicas de anestesia loco-regional ahorradoras de opioides, optimizar las estrategias de prevención de náuseas y vómitos, o prever un monitoreo más prolongado en recuperación. A la inversa, una predicción de recuperación excelente podría sustentar protocolos de alta temprana en conjunto con los criterios clínicos de alta establecidos o cirugías ambulatorias. Asimismo, permite informar proactivamente al paciente sobre sus expectativas de recuperación de manera cuantitativa.

El presente estudio presenta algunas limitaciones que deben considerarse. Aunque este trabajo calibró el modelo utilizando los mismos 318 pacientes de la fase de validación inicial, sometimos el marco a pruebas de validación interna (validación cruzada y 500 ciclos de *bootstrapping*) que no evidenciaron sobreajuste (optimismo de -0.007 en el Factor K y de 0.035 en el Intercepto b0). Los resultados de la simulación

en “pacientes no vistos” (R^2 OOS=0.915) fueron superponibles a los del modelo original. Si bien el origen bicéntrico de la muestra (Buenos Aires y Corrientes) y las exhaustivas pruebas de validación interna respaldan la estabilidad empírica del modelo, se debe enfatizar que la herramienta permanece en una etapa de desarrollo. La validación prospectiva en cohortes externas independientes constituirá un paso clave y un requisito indispensable para consolidar su generalizabilidad, antes de poder recomendar su utilización clínica rutinaria.

En conclusión, el marco metodológico MIMIC-CPM se presenta como una aproximación estadística factible y potencialmente útil, a nivel local, para estimar de forma individualizada la calidad de recuperación post-anestésica. La convergencia estructural-empírica observada ($\Delta = 0.309$), la concordancia entre los distintos motores de estimación y la ausencia de evidencia de sobreajuste en la validación interna sugieren una adecuada estabilidad del modelo en la población de derivación. El dashboard interactivo derivado de la validación al español rioplatense del QoR-15 podría facilitar su evaluación en la práctica clínica; no obstante, su utilización rutinaria requiere una validación externa prospectiva previa en cohortes independientes.

Conflicto de intereses: Ninguno para declarar

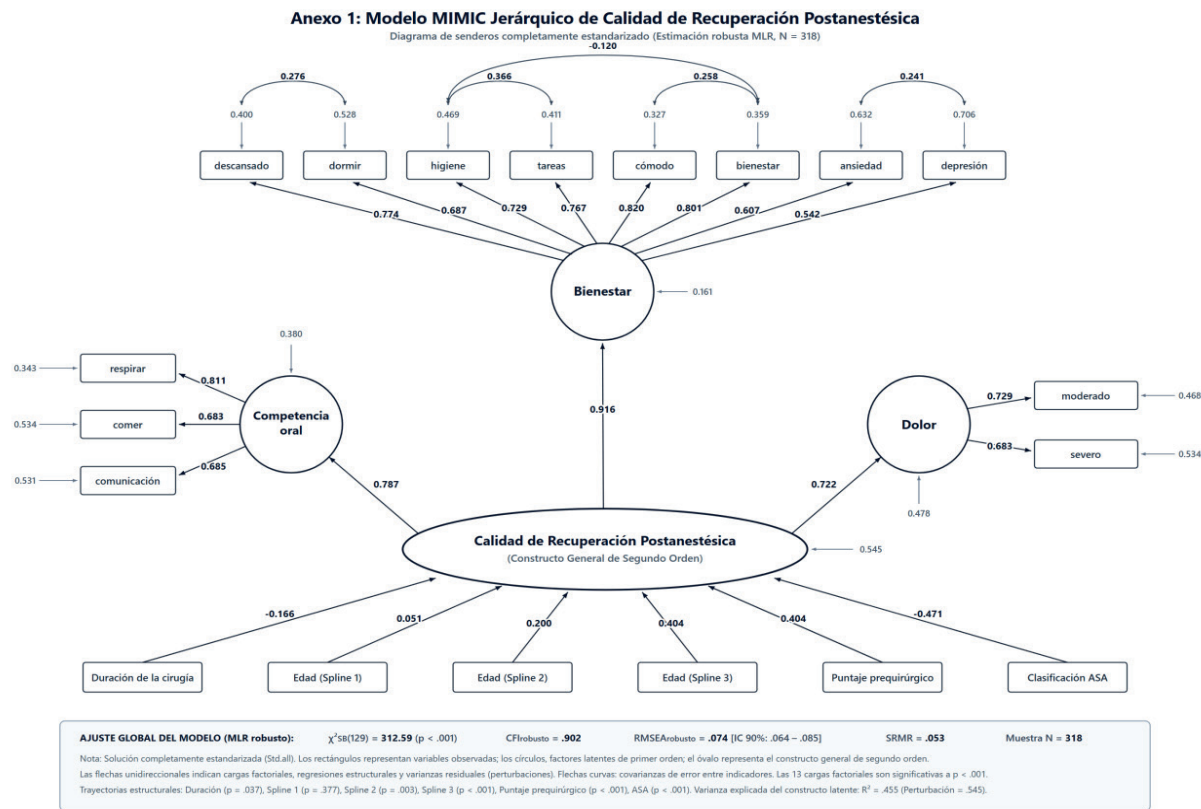
Bibliografía

1. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* 2010; 21:128-38.
2. Stark PA, Myles PS, Burke JA. Development and psychometric evaluation of a postoperative quality of recovery score: The QoR-15. *Anesthesiology* 2013; 118: 1332-40.
3. Hair J, Black W, Babin B, Anderson R. Multivariate Data Analysis. 8th Edition. Cengage Learning EMEA; 2018.
4. Grande Ratti MF, Pérez Manelli RY, Vázquez Peña FR, Cordon Pozo E. Modelos MIMIC: de la neurociencia a las ciencias de la salud. *Rev Hosp Ital B Aires* 2024; 44: e0000344.
5. Benítez SM, Puyol RB, Grande Ratti MF, Vázquez Navarro M, Vázquez Peña FR, Cordon Pozo E. Adaptación y validación del Cuestionario QoR-15 al español rioplatense. *Medicina (B Aires)* xxxx; xx: xx-xx.
6. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Cham: Springer; 2019.
7. Rosseel Y, Loh WW. A structural after measurement approach to structural equation modeling. *Psychol Methods*. 2024.
8. Gorsuch RL. Factor Analysis. Classic 2nd ed. New York: Routledge; 2015.
9. DiStefano C, Zhu M, Mándrilă D. Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Pract Assess Res Eval* 2009;14: 1-11.
10. Hogg RV, Tanis EA, Zimmerman DL. Probability and Statistical Inference. 9th ed. Boston: Pearson; 2015.

11. Hogg RV, Craig AT. Introduction to Mathematical Statistics. 4th ed. New York: Macmillan; 1978.
12. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. Cham: Springer; 2015.
13. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; 385: e078378.
14. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15: 361-87.
15. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I - continuous outcomes. *Stat Med* 2019; 38: 1262-75.
16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the achilles heel of predictive analytics. *BMC Med* 2019; 17: 230.
17. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

Anexo 1: Modelo MIMIC Jerárquico de Calidad de Recuperación Pos-anestésica

Estructura latente de segundo orden sustentada en tres factores de primer orden (Bienestar, Competencia oral, Dolor) derivados de los 13 ítems observados, con regresión estructural sobre la clasificación ASA, edad en *splines* (df = 3), duración quirúrgica lineal y puntaje prequirúrgico basal.



Anexo 2: Formulario del Cuestionario QoR-13 (Versión Adaptada al Español Rioplatense)

Versión Validada de 13 ítems

ANEXO N° 2 - FORMULARIO QoR-13 (VERSIÓN ADAPTADA EN ESPAÑOL RIOPLATENSE)

Quality of Recovery-13 (QoR-13) - Versión Depurada en Español Rioplatense

Fecha: Estudio ID: ☐ Prequirúrgico ☐ Posquirúrgico

PARTE A: ¿CÓMO SE HA SENTIDO EN LAS ÚLTIMAS 24 HORAS?

Respuestas puntuadas de 0 a 10 (0 = en ningún momento [mal], 10 = en todo momento [excelente])

1. Capaz de respirar sin problemas

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

2. ¿Puede disfrutar la comida?

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

3. ¿Se siente descansado?

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

4. ¿Ha dormido bien?

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

5. ¿Puede cuidar su higiene personal sin requerir ayuda?

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

6. ¿Capaz de comunicarse correctamente con familia o amigos?

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

7. ¿Puede regresar al trabajo y a las tareas cotidianas?

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

8. Se siente cómodo y en control

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

9. Tiene sensación de bienestar general

En ningún momento 0 1 2 3 4 5 6 7 8 9 10 En todo momento

PARTE B: ¿HA EXPERIMENTADO ALGUNO DE LOS SIGUIENTES ESTADOS EN LAS ÚLTIMAS 24 HORAS?

Respuestas puntuadas de 10 a 0 (10 = en ningún momento [excelente], 0 = en todo momento [mal])

10. Dolor moderado

En ningún momento 10 9 8 7 6 5 4 3 2 1 0 En todo momento

11. Dolor severo

En ningún momento 10 9 8 7 6 5 4 3 2 1 0 En todo momento

12. Preocupación o ansiedad

En ningún momento 10 9 8 7 6 5 4 3 2 1 0 En todo momento

13. Tristeza o depresión

En ningún momento 10 9 8 7 6 5 4 3 2 1 0 En todo momento

DATOS DEMOGRÁFICOS

Edad Sexo
 ☐ F
 ☐ M

Estado Físico ASA: _____

COMORBILIDADES

- ☐ Cardiovascular
☐ Renal
☐ Respiratoria
☐ Gastrointestinal
☐ Neurológica
☐ Endocrinológica
☐ Otras

TIPO DE CIRUGÍA

- ☐ Programada
☐ No programada
☐ Cardiororácica
☐ General
☐ Ortopédica
☐ Neurocirugía
☐ Vascular
☐ Maxilofacial
☐ Urológica

Duración cirugía: _____ Min

Evaluación post: _____ horas