

Práctica Empresarial: Segmentación de diagnósticos y análisis financiero de pacientes en el servicio de hospitalización mediante modelos de clustering y predictivos de ingreso por paciente.

Keren Slendy Rodríguez Calderón

Trabajo de grado para optar al título de Matemática

Director

Andrés Sebastián Ríos Gutiérrez

Doctor en Ciencias-Estadística

Codirector

Jorge Villamizar Morales

Doctor en Ciencias Aplicadas

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Matemáticas

Bucaramanga

2026

Dedicatoria

A mi abuelita Gloria, aunque ya no pueda abrazarte, te llevo conmigo en cada latido de mi corazón. Tu amor me acompaña y tu recuerdo me inspira siempre.

Agradecimientos

A Dios, por el privilegio de la vida, por acompañarme en cada etapa de este camino y por darme la fortaleza, sabiduría y perseverancia necesarias para culminar esta meta. Reconozco que cada aprendizaje, oportunidad y logro alcanzado han sido posibles gracias a su amor, su gracia y su infinita misericordia.

A mis padres, por su amor incondicional, sus sacrificios y su apoyo constante. Gracias por creer en mí, por sostenerme en los momentos difíciles y por enseñarme con su ejemplo el valor del esfuerzo, la responsabilidad y la perseverancia. Este logro también es de ustedes.

A mis profesores, por los conocimientos y enseñanzas compartidas durante mi formación académica. De manera especial, agradezco a mi director de práctica, el profesor Andrés Ríos, por su guía, paciencia, dedicación y acompañamiento constante. Sus orientaciones, su calidad humana y su confianza fueron fundamentales para el desarrollo de este trabajo.

Asimismo, expreso mi sincero agradecimiento a mi codirector, el profesor Jorge Villamizar, por su apoyo permanente, sus consejos y la confianza depositada en mis capacidades. Su acompañamiento representó una contribución valiosa en mi formación académica y personal.

A la Fundación Oftalmológica de Santander – FOSCAL, por permitirme desarrollar esta práctica profesional en un entorno de aprendizaje enriquecedor. Agradezco especialmente a la doctora Liliana Landazábal por su liderazgo, confianza, orientación y compromiso durante el desarrollo del proyecto. Su apoyo fue esencial para culminar satisfactoriamente esta experiencia.

También agradezco a mis amigos, compañeros y a todas las personas que hicieron parte de mi paso por la universidad. Gracias por las palabras de ánimo, la ayuda generosa, las conversaciones, las risas compartidas y los gestos de apoyo que hicieron este camino más llevadero y significativo.

Este logro no es únicamente mío; también pertenece a quienes me acompañaron, confiaron en mí y me recordaron que podía seguir adelante aun en los momentos de mayor dificultad. Con fe

y gratitud, encomiendo a Dios los frutos de este trabajo y confío en que las enseñanzas adquiridas serán la base de nuevos caminos y oportunidades.

Este trabajo de grado de pregrado se enmarca dentro de los productos del Grupo de Investigación en Procesos Estocásticos de la Universidad Nacional de Colombia sede Bogotá, registrado en el GrupLAC de MinCiencias. Corresponde a un producto dirigido por el profesor Andrés Ríos, en el contexto de su Beca de Excelencia Doctoral del Bicentenario Corte I, perteneciente a la convocatoria del Fondo de Ciencia, Tecnología e Innovación del Sistema General de Regalías para la conformación de priorizados y aprobados por el OCAD.

Tabla de Contenido

	Pág.
Introducción	20
1. Generalidades	23
1.1 Planteamiento del problema	23
1.2 Justificación	24
1.2.1 Necesidad institucional	25
1.2.2 Pertinencia e impacto esperado	26
1.2.3 Innovación y aporte del proyecto	27
2. Objetivos	29
2.1 Trazabilidad administrativa del proyecto	29
2.2 Objetivo general	29
2.3 Objetivos específicos	30
3. Marco referencial, estado del arte y consideraciones éticas	30
3.1 Marco teórico	30
3.2 Métodos alternativos considerados	32
3.3 Componente financiero: del episodio clínico a la caracterización económica	33
3.4 Estado del arte	33
3.5 Consideraciones éticas, privacidad y uso secundario de datos	34
4. Metodología	35
4.1 Diseño, contexto empresarial y alcance de la práctica	35
4.2 Datos, cohorte de análisis y flujo metodológico	36
4.3 Preparación de variables, recodificación y análisis bivariado preliminar	37

4.4	Reclasificación clínica y agrupamiento no supervisado	40
4.5	Caracterización clínica, análisis financiero y residuos ajustados	41
4.6	Reproducibilidad computacional y trazabilidad del código	43
5.	Preparación de datos y análisis preliminar	45
5.1	Estructura general, variables analíticas e integridad de la base	45
5.2	Exploración preliminar, concentración diagnóstica y recodificación categórica	47
5.3	Definición de la cohorte 2024 sin COVID-19	49
5.4	Sensibilidad de recodificaciones y análisis bivariado global	50
5.5	Preparación preliminar de variables para el agrupamiento clínico	54
5.6	Transformación financiera preliminar	58
6.	Análisis post hoc de asociaciones	60
6.1	Método, criterios y alcance del análisis post hoc	60
6.2	Resultados por cruce priorizado	62
6.2.1	Servicio y diagnóstico de egreso Top 50	62
6.2.2	Servicio y fallecimiento	64
6.2.3	Servicio y asignación de enfermería Top 12	65
6.2.4	Asignación de enfermería Top 12 y EPS Top 12	66
6.2.4.1	Nota sobre ceros observados y restricciones estructurales.	67
6.3	Síntesis interpretativa y alcance del análisis post hoc	68
7.	Agrupamiento clínico no supervisado	70
7.1	Propósito, unidad de análisis y cohorte	70
7.1.1	Selección de la cohorte de análisis	73
7.2	Reclasificación clínica, discrepancia y sensibilidad COVID-19	75
7.2.1	Discrepancia entre clasificación original y reconstruida	79
7.2.2	Sensibilidad por exclusión de COVID-19	80
7.3	Preparación de variables y estrategia de agrupamiento	81

7.3.1	Estrategia de agrupamiento	83
7.4	Selección del número de grupos y solución final	85
7.5	Caracterización de grupos y análisis de mortalidad	87
7.5.1	Mortalidad: análisis crudo, pruebas de comparación y ajuste por edad/sexo	91
7.6	Estabilidad, reproducibilidad, calidad del dato y controles técnicos	94
7.6.1	Calidad del dato y controles técnicos	96
7.7	Base médica final, limitaciones y conclusión	98
7.7.1	Limitaciones metodológicas	99
7.7.2	Conclusión del capítulo	101
8.	Análisis financiero por grupo clínico	102
8.1	Objetivo, alcance y exploración cuantitativa inicial	102
8.2	Transformación de la respuesta financiera y preparación de la estancia	104
8.3	Diseño ANCOVA, supuestos evaluados y diagnóstico del modelo	107
8.4	Resultados inferenciales y estimaciones marginales ajustadas por estancia real	109
8.5	Concentración del ingreso y análisis focalizado del grupo no transmisible C03	113
8.6	Residuo financiero ajustado y priorización exploratoria	116
8.6.0.1	Sensibilidad de la regla exploratoria P75/P90.	117
8.7	Síntesis, limitaciones e implicaciones de gestión	118
9.	Conclusiones generales, limitaciones y recomendaciones	119
9.1	Conclusiones alineadas con los objetivos específicos	119
9.2	Limitaciones del estudio	121
9.3	Recomendaciones	122
	Referencias Bibliográficas	124
	Apéndices	130

Lista de Tablas

Tabla 1.	Síntesis del marco teórico y relación con el análisis aplicado	31
Tabla 2.	Flujo metodológico general de datos	37
Tabla 3.	Variables disponibles/preparadas y uso metodológico	38
Tabla 4.	Pseudocódigo general del flujo computacional utilizado	44
Tabla 5.	Resumen de reproducibilidad computacional del análisis	45
Tabla 6.	Dimensiones analíticas y variables objetivo iniciales	46
Tabla 7.	Recodificación de variables categóricas por frecuencia acumulada . . .	48
Tabla 8.	Comparación de cohortes de episodios hospitalarios 2024 según exclusión COVID-19	49
Tabla 9.	Sensibilidad de la exclusión COVID-19 sobre la composición relativa de macrogrupos	50
Tabla 10.	Análisis de sensibilidad y configuración de variables categóricas	51
Tabla 11.	Indicadores globales del análisis bivariado final	52
Tabla 12.	Asociaciones con mayor V de Cramér e intervalos bootstrap disponibles	53
Tabla 13.	Cruces no interpretados como hallazgos sustantivos por dependencia con- ceptual	54
Tabla 14.	Distribución de episodios por capítulo CIE-10	55
Tabla 15.	Análisis de sensibilidad del número de componentes SVD en el macro- grupo no transmisible	56
Tabla 16.	Análisis de sensibilidad del número de componentes SVD en el macro- grupo infecciosa no COVID	56

Tabla 17.	Métricas internas preliminares para evaluación de k en macrogrupos subdivididos	57
Tabla 18.	Comparación de transformaciones candidatas del ingreso total	59
Tabla 19.	Cruces priorizados para el análisis post hoc	61
Tabla 20.	Resumen global de los cruces post hoc	62
Tabla 21.	Celdas con mayor contribución al chi-cuadrado: servicio y diagnóstico de egreso Top 50	63
Tabla 22.	Celdas con mayor contribución al chi-cuadrado: servicio y fallecimiento	64
Tabla 23.	Celdas con mayor contribución al chi-cuadrado: servicio y asignación de enfermería Top 12	65
Tabla 24.	Celdas con mayor contribución al chi-cuadrado: asignación de enfermería Top 12 y EPS Top 12	66
Tabla 25.	Principales celdas con cero observado y frecuencia esperada positiva en servicio y diagnóstico Top 50	68
Tabla 26.	Clasificación interpretativa de los cruces post hoc	69
Tabla 27.	Definiciones operativas utilizadas en el capítulo	72
Tabla 28.	Verificación de unicidad del identificador de episodio en la base final . .	73
Tabla 29.	Distribución de episodios por año de análisis	74
Tabla 30.	Orden jerárquico de reclasificación clínica	76
Tabla 31.	Códigos y capítulos CIE-10 usados en la reclasificación clínica	77
Tabla 32.	Patrones textuales principales usados en la reclasificación clínica	78
Tabla 33.	Verificación de asignación completa posterior a la reclasificación	79
Tabla 34.	Discrepancia clínica por macrogrupo en la base final sin COVID-19 . .	80
Tabla 35.	Sensibilidad descriptiva por exclusión de COVID-19	81
Tabla 36.	Variables disponibles/preparadas por macrogrupo para el agrupamiento y la caracterización	82

Tabla 37.	Resumen de codificación y reducción SVD en macrogrupos subdivididos	83
Tabla 38.	Parámetros computacionales de la estrategia de agrupamiento	84
Tabla 39.	Evaluación de k en no transmisible	86
Tabla 40.	Evaluación de k en infecciosa no COVID	86
Tabla 41.	Nombres interpretativos de los grupos clínicos finales	88
Tabla 42.	Mortalidad cruda por grupo clínico final con intervalos de confianza del 95 %	89
Tabla 43.	Caracterización de grupos priorizados: mortalidad y volumen	90
Tabla 44.	Caracterización de grupos priorizados: distribución de edad	90
Tabla 45.	Comparación de mortalidad cruda de cada grupo frente al resto de la cohorte	92
Tabla 46.	Mortalidad ajustada por edad y sexo mediante estandarización marginal	93
Tabla 47.	Estabilidad de la solución mediante ARI en 30 corridas	95
Tabla 48.	Estabilidad de la solución mediante NMI en 30 corridas	95
Tabla 49.	Datos faltantes en variables clave de la base final	97
Tabla 50.	Utilidad exploratoria de los grupos finales frente a macrogrupos para explicar mortalidad	99
Tabla 51.	Descripción inicial de las variables cuantitativas del análisis financiero .	103
Tabla 52.	Comparación metodológica de transformaciones de la respuesta financiera	106
Tabla 53.	Comparación de modelos sobre la respuesta logarítmica robusta	108
Tabla 54.	Pruebas globales del modelo ANCOVA con interacción	110
Tabla 55.	Comparaciones post hoc seleccionadas entre grupos clínicos	111
Tabla 56.	Estimaciones marginales ajustadas por grupo clínico y nivel real de estancia	112
Tabla 57.	Contribución financiera acumulada por grupo clínico	114
Tabla 58.	Subgrupos clínicos con mayor ingreso dentro de No transmisible C03 .	115
Tabla 59.	Pruebas de Kruskal–Wallis sobre el residuo financiero ajustado	117

Tabla 60.	Priorización exploratoria de episodios según residuo financiero positivo ajustado	118
Tabla 61.	Principales diagnósticos en la categoría mal definidas e inespecíficas . .	149
Tabla 62.	Matriz de reclasificación: categorías reconstruidas COVID-19, infecciosas, lesiones y mal definidas	150
Tabla 63.	Matriz de reclasificación: categorías reconstruidas materna, no transmisible y perinatal	151
Tabla 64.	Sensibilidad SVD en no transmisible	151
Tabla 65.	Sensibilidad SVD en infecciosa no COVID	152
Tabla 66.	Evaluación resumida de macrogrupos conservados como grupos únicos	152
Tabla 67.	Criterio de conservación como grupo único	153
Tabla 68.	Resumen de estabilidad mediante ARI con 30 corridas	153
Tabla 69.	Resumen de estabilidad mediante NMI con 30 corridas	154
Tabla 70.	Mortalidad cruda por macrogrupo con IC95 %	154
Tabla 71.	Comparación de mortalidad por macrogrupo frente al resto de la cohorte	155
Tabla 72.	Mortalidad ajustada por edad y sexo por grupo clínico final	156
Tabla 73.	Modelo logístico ajustado por edad y sexo	157
Tabla 74.	Mortalidad según presencia de discrepancia clínica	158
Tabla 75.	Discrepancia clínica por macrogrupo	158
Tabla 76.	Discrepancia clínica por grupo clínico final	159
Tabla 77.	Etiquetas interpretativas de los subgrupos clínicos finales	161
Tabla 78.	Criterio principal utilizado para nombrar cada grupo	162
Tabla 79.	Perfil cuantitativo de grupos priorizados	163
Tabla 80.	Atributos dominantes de grupos priorizados	164
Tabla 81.	Tendencia mensual de indicadores de calidad del dato durante 2024 . .	179
Tabla 82.	Datos faltantes en variables clave de la base final	180
Tabla 83.	Revisión de plausibilidad clínica básica	180

Tabla 84.	Resumen de valores atípicos por regla IQR	181
Tabla 85.	Resumen de diagnósticos raros por nombre diagnóstico	181
Tabla 86.	Verificación de unicidad del identificador de episodio en la base final . .	181
Tabla 87.	Información no disponible para análisis adicionales solicitados o sugeridos	182
Tabla 88.	Resumen del error de agregación microgrupos-centroides en el espacio SVD	183
Tabla 89.	Sensibilidad de la solución al excluir CIE-3 con frecuencia menor o igual a 2	183
Tabla 90.	Reproducibilidad de métricas internas en 30 corridas	184
Tabla 91.	Comparación exploratoria con métodos alternativos sobre el espacio SVD	184
Tabla 92.	Comparación de modelos de mortalidad con macrogrupo frente a grupo clínico final	185
Tabla 93.	Diagnóstico comparativo de transformaciones de la respuesta financiera	192
Tabla 94.	Episodios con mayor distancia de Cook en el ANCOVA con interacción	193
Tabla 95.	Estimaciones marginales ajustadas completas con intervalos de confianza aproximados	194
Tabla 96.	Diagnósticos con mayor ingreso acumulado en No transmisible C03 . .	199
Tabla 97.	Diagnósticos con mayor variabilidad financiera en No transmisible C03	201
Tabla 98.	Sensibilidad de la clasificación de residuo financiero positivo ante cam- bios de percentil	204
Tabla 99.	Concordancia de clasificaciones alternativas frente a P75/P90	204
Tabla 100.	Información externa necesaria para validación futura	207

Lista de Figuras

Figura 1.	Mapa de calor de los tamaños de efecto medidos mediante V de Cramér.	52
Figura 2.	Relación exploratoria entre estancia y respuesta financiera transformada por grupo clínico.	104
Figura 3.	Residuos frente a valores ajustados del ANCOVA con interacción.	109
Figura 4.	Estimaciones marginales ajustadas por grupo clínico y nivel real de estancia.	113
Figura 5.	Participación del ingreso acumulado por grupo clínico.	114
Figura 6.	Aprobación del Comité de Ética de FOSCAL para el proyecto.	130
Figura 7.	Porcentaje de valores faltantes por variable en la base analítica 2021–2024.	132
Figura 8.	Cardinalidad de las variables disponibles para el análisis preliminar.	133
Figura 9.	Distribución de la edad de los episodios hospitalarios.	133
Figura 10.	Distribución de la estancia hospitalaria y presencia de valores extremos.	134
Figura 11.	Concentración acumulada de los diagnósticos de egreso.	134
Figura 12.	Sensibilidad de la composición de macrogrupos clínicos con y sin episodios COVID-19.	135
Figura 13.	Distribución de las categorías principales del diagnóstico de egreso después de la recodificación Top 50.	136
Figura 14.	Distribución de la variable grupo de edad utilizada en el proceso de agrupamiento.	137
Figura 15.	Distribución de los principales prefijos CIE-3 utilizados en el proceso de agrupamiento.	137
Figura 16.	Curva de varianza explicada acumulada por número de componentes SVD en el macrogrupo no transmisible.	138

Figura 17.	Curva de varianza explicada acumulada por número de componentes SVD en el macrogrupo infecciosa no COVID.	139
Figura 18.	Coeficiente de silueta según el número de grupos en el macrogrupo no transmisible.	139
Figura 19.	Índice Calinski–Harabasz según el número de grupos en el macrogrupo no transmisible.	140
Figura 20.	Índice Davies–Bouldin según el número de grupos en el macrogrupo no transmisible.	140
Figura 21.	Coeficiente de silueta según el número de grupos en el macrogrupo infecciosa no COVID.	141
Figura 22.	Índice Calinski–Harabasz según el número de grupos en el macrogrupo infecciosa no COVID.	141
Figura 23.	Índice Davies–Bouldin según el número de grupos en el macrogrupo infecciosa no COVID.	142
Figura 24.	Estabilidad de la solución de agrupamiento mediante índice de Rand ajustado bajo cambios de semilla y remuestreo.	142
Figura 25.	Estabilidad de la solución de agrupamiento mediante información mutua normalizada bajo cambios de semilla y remuestreo.	143
Figura 26.	Distribución del ingreso total original antes de aplicar transformaciones. .	143
Figura 27.	Comparación de dispersión y valores extremos entre transformaciones candidatas del ingreso total.	144
Figura 28.	Gráfico Q–Q de la transformación logarítmica robusta del ingreso total. .	145
Figura 29.	Contribución porcentual al chi-cuadrado en el cruce Servicio y diagnóstico de egreso Top 50.	146
Figura 30.	Contribución porcentual al chi-cuadrado en el cruce Servicio y fallecimiento.	146

Figura 31.	Contribución porcentual al chi-cuadrado en el cruce Servicio y asignación de enfermería Top 12.	147
Figura 32.	Contribución porcentual al chi-cuadrado en el cruce Asignación de enfermería Top 12 y EPS Top 12.	147
Figura 33.	Composición clínica de la cohorte 2024 según clasificación reconstruida .	148
Figura 34.	Distribución por servicio dentro de cada grupo clínico final	164
Figura 35.	Distribución por grupo de edad dentro de cada grupo clínico final	166
Figura 36.	Distribución por unidad médica dentro de cada grupo clínico final	168
Figura 37.	Distribución por sexo dentro de cada grupo clínico final	170
Figura 38.	Diagnósticos de egreso más frecuentes dentro de cada grupo clínico final	172
Figura 39.	Mortalidad intragrupo por grupo clínico final	173
Figura 40.	Aporte global de cada grupo clínico final a las muertes	175
Figura 41.	Comparación entre mortalidad intragrupo y aporte global a muertes . . .	176
Figura 42.	Comparación de indicadores de calidad del dato entre 2021–2023 y 2024	178
Figura 43.	Tendencia anual de indicadores de calidad del dato	179
Figura 44.	Histograma del ingreso total truncado visualmente en el percentil 99. . .	186
Figura 45.	Q-Q plot del ingreso total original.	187
Figura 46.	Q-Q plot de log(Ingreso Total).	188
Figura 47.	Boxplot global del ingreso total original.	189
Figura 48.	Boxplot global de log(Ingreso Total).	189
Figura 49.	Distribución de la respuesta financiera transformada por grupo clínico. . .	190
Figura 50.	Relación global entre estancia real e ingreso transformado.	191
Figura 51.	Relación entre estancia real e ingreso transformado por grupo clínico. . .	191
Figura 52.	Q–Q plot de residuos del ANCOVA con interacción.	193
Figura 53.	Distancia de Cook del ANCOVA con interacción.	193
Figura 54.	Ingreso acumulado por subgrupo clínico en No transmisible C03.	196
Figura 55.	Diagnósticos con mayor ingreso acumulado en No transmisible C03. . . .	197

Figura 56.	Diagnósticos con mayor ingreso mediano en No transmisible C03.	197
Figura 57.	Relación entre estancia e ingreso en diagnósticos prioritarios de No transmisible C03.	198
Figura 58.	Diagnósticos con mayor variabilidad financiera en No transmisible C03. .	200
Figura 59.	Diagnósticos con mayor variabilidad de estancia en No transmisible C03.	201
Figura 60.	Residuo financiero ajustado por servicio.	202
Figura 61.	Residuo financiero ajustado por unidad médica.	202
Figura 62.	Residuo financiero ajustado por EPS.	203
Figura 63.	Residuo financiero ajustado por clase de movimiento de alta.	203
Figura 64.	Diagnósticos asociados a residuo financiero positivo muy alto.	206

Lista de Apéndices

Apéndice A. Aprobación del Comité de Ética de FOSCAL	130
Apéndice B. Trazabilidad administrativa complementaria	130
Apéndice C. Complementos del Capítulo 5	131
Apéndice D. Complementos del Capítulo 6	145
Apéndice E. Complementos del Capítulo 7	148
Apéndice F. Complementos del Capítulo 8	185

Resumen

Título: Práctica Empresarial: Segmentación de diagnósticos y análisis financiero de pacientes en el servicio de hospitalización mediante modelos de clustering y predictivos de ingreso por paciente.

Autor: Keren Slendy Rodríguez Calderón

Palabras clave: agrupamiento no supervisado, episodios hospitalarios, ingreso por episodio, análisis financiero, CIE-10, FOSCAL.

Descripción:

Esta práctica empresarial desarrolló una ruta analítica para segmentar episodios hospitalarios y caracterizar su comportamiento financiero en el servicio de hospitalización de la Fundación Oftalmológica de Santander – Clínica FOSCAL. Se utilizaron registros secundarios anonimizados de 2021–2024, con análisis en la cohorte 2024 sin episodios COVID-19, tomando el episodio hospitalario como unidad de análisis. La metodología incluyó depuración, normalización, control de duplicados, evaluación de faltantes, recodificación y reclasificación clínica mediante códigos CIE-10 y nombres diagnósticos. Para evitar circularidad, el agrupamiento se construyó con variables diagnósticas y demográficas, excluyendo mortalidad, ingreso, estancia, EPS, servicio y unidad médica. Se combinaron codificación *one-hot*, *TruncatedSVD*, *MiniBatchKMeans* y agrupamiento jerárquico de Ward. La solución final definió diez grupos clínico-diagnósticos, caracterizados por mortalidad, estancia, composición diagnóstica e ingreso. El análisis financiero empleó el ingreso total como variable, sin interpretarlo como costo, utilidad ni rentabilidad, y aplicó transformación logarítmica robusta, ANCOVA con interacción grupo clínico–estancia, errores estándar HC3, comparaciones ajustadas y residuos financieros. Los resultados evidenciaron heterogeneidad clínica y financiera, concentración del ingreso en grupos no transmisibles y episodios con ingreso superior al esperado. El estudio aporta una herramienta reproducible para priorización institucional, auditoría financiera y calidad del dato, sin establecer causalidad ni reemplazar la auditoría clínica individual.

* Trabajo de grado.

** Facultad de Ciencias. Escuela de Matemáticas. Director: Andrés Sebastián Ríos Gutiérrez. Codirector: Jorge Villamizar Morales

Abstract

Title: Business Internship: Diagnostic Segmentation and Financial Analysis of Patients in the Inpatient Service Using Clustering and Patient-Level Revenue Prediction Models

Author: Keren Slendy Rodríguez Calderón

Key Words: unsupervised clustering, hospital episodes, income per episode, financial analysis, ICD-10, FOSCAL.

Description:

This business internship developed an analytical workflow to segment hospital episodes and characterize their financial behavior in the hospitalization service of Fundación Oftalmológica de Santander – Clínica FOSCAL. Anonymized secondary records from 2021–2024 were analyzed, focusing on the 2024 cohort after excluding COVID-19 episodes and using the hospital episode as the unit of analysis. The methodology included data cleaning, normalization, duplicate control, missing-value assessment, category recoding, and clinical reclassification based on ICD-10 codes and diagnostic names. To avoid circularity, clustering used diagnostic and demographic variables while excluding mortality, income, length of stay, insurer, service, and medical unit. The strategy combined one-hot encoding, TruncatedSVD, MiniBatchKMeans, and Ward hierarchical clustering. Ten clinical-diagnostic groups were identified and characterized by mortality, length of stay, diagnostic composition, and income. Financial analysis used total income without interpreting it as cost, profit, or profitability, applying robust logarithmic transformation, ANCOVA with clinical group–length of stay interaction, HC3 standard errors, adjusted comparisons, and financial residuals. Results showed clinical and financial heterogeneity, concentration of cumulative income in noncommunicable groups, and episodes with income above expected values. The study provides a reproducible tool for institutional prioritization, financial auditing, and data quality review without establishing causality or replacing individual clinical auditing and oversight.

* Bachelor Thesis.

** Facultad de Ciencias. Escuela de Matemáticas. Director: Andrés Sebastián Ríos Gutiérrez. Codirector: Jorge Villamizar Morales

Introducción

Una de las preguntas que con frecuencia se plantea frente a la formación matemática es cuál puede ser el aporte de un matemático en equipos financieros o en procesos de análisis de datos médicos. Esta inquietud surge porque, en muchos contextos, la labor del matemático se asocia principalmente con la docencia o con el desarrollo teórico. No obstante, la formación matemática proporciona herramientas rigurosas para estructurar, modelar e interpretar fenómenos complejos, especialmente en escenarios donde la toma de decisiones depende del análisis sistemático de grandes volúmenes de información.

En el sector salud, las instituciones generan de manera permanente datos clínicos, administrativos, operativos y financieros. Cada episodio de atención, diagnóstico, ingreso hospitalario, egreso, procedimiento, medicamento suministrado o recurso facturado constituye un registro que puede aportar información relevante para la gestión institucional. Sin embargo, la disponibilidad de datos no implica, por sí misma, la generación de conocimiento. Para que esta información sea útil, debe ser organizada, depurada, transformada y analizada mediante métodos que permitan identificar patrones, relaciones y comportamientos que no siempre son evidentes a partir de indicadores agregados.

En este contexto, la participación de un matemático adquiere especial relevancia. Su formación permite abordar los datos no solo como reportes descriptivos, sino como estructuras susceptibles de ser estudiadas mediante modelos, algoritmos y criterios analíticos. En equipos financieros y clínicos, el matemático puede contribuir a comprender la variabilidad de los ingresos, detectar comportamientos atípicos, estudiar relaciones entre variables operativas y económicas, construir indicadores ajustados, reconocer perfiles de pacientes y desarrollar metodologías reproducibles que apoyen la toma de decisiones institucional.

La presente práctica empresarial se desarrolló en la Fundación Oftalmológica de Santander

– Clínica FOSCAL, institución que dispone de un volumen considerable de información derivada de sus procesos asistenciales y administrativos. Esta experiencia fue posible gracias a la gestión y acompañamiento del Doctor en Ciencias Aplicadas Jorge Villamizar, así como a la orientación del Doctor en Estadística Andrés Ríos. Enmarcada en el Programa de Matemáticas de la Universidad Industrial de Santander, la práctica constituyó una oportunidad para articular la formación matemática con una necesidad real del sector salud, integrando análisis de datos, finanzas, operación hospitalaria e interpretación clínica.

La motivación principal del proyecto radicó en evidenciar el potencial de la matemática en escenarios aplicados, especialmente al tratarse de una de las primeras experiencias de este tipo impulsadas desde la Escuela de Matemáticas. Más allá del cumplimiento de un requisito académico, la práctica buscó mostrar que las herramientas matemáticas pueden contribuir a la comprensión y solución de problemas institucionales reales. El entorno hospitalario ofreció un contexto especialmente pertinente debido a la complejidad de sus procesos, la heterogeneidad de los pacientes atendidos y la necesidad de comprender mejor el comportamiento clínico y financiero del servicio de hospitalización.

Durante el desarrollo de la práctica se trabajó con bases de datos anonimizadas de la Clínica FOSCAL correspondientes al periodo 2021–2024. Estas bases incluían información sobre episodios de hospitalización, diagnósticos de egreso, variables demográficas, servicios de atención, unidades médicas, estancia hospitalaria e ingresos asociados a distintos componentes financieros. A partir de esta información, se planteó una estrategia de análisis orientada a transformar registros rutinarios en insumos útiles para la gestión institucional, conectando la información clínica con el comportamiento financiero de los episodios hospitalarios.

El análisis realizado partió del reconocimiento de que los datos hospitalarios no pueden interpretarse de manera aislada del contexto en el que se producen. Un diagnóstico, una estancia prolongada, un ingreso elevado o una unidad de atención específica reflejan procesos clínicos, decisiones operativas, rutas asistenciales y condiciones administrativas. Por esta razón, el proyecto integró herramientas matemáticas, estadísticas y computacionales con una interpretación cuidadosa

del entorno institucional, con el fin de evitar conclusiones exclusivamente numéricas y avanzar hacia resultados comprensibles y útiles para los equipos clínicos y financieros.

El proyecto se estructuró en dos componentes principales. El primero correspondió a la segmentación clínica de episodios hospitalarios mediante técnicas de aprendizaje no supervisado, particularmente métodos de agrupamiento. Esta etapa permitió identificar grupos de episodios con características clínicas y diagnósticas relativamente homogéneas, facilitando una lectura por perfiles y no únicamente una descripción general del servicio de hospitalización. Los resultados permitieron reconocer diferencias relevantes entre grupos en términos clínicos, operativos y de mortalidad.

El segundo componente correspondió al análisis financiero de los grupos clínicos obtenidos. Una vez construida la segmentación, se integraron variables relacionadas con ingreso por episodio, estancia hospitalaria y componentes financieros asociados a servicios, materiales, medicamentos y honorarios profesionales. El propósito fue describir las diferencias financieras entre grupos clínicos e identificar episodios con ingresos superiores a lo esperado, considerando elementos como el diagnóstico y la duración de la estancia. Este enfoque permitió abordar una pregunta institucional relevante: qué factores descriptivos se asocian con comportamientos financieros distintos en episodios con diagnósticos similares.

Es importante señalar que el análisis financiero se centró en el ingreso y no en la utilidad o rentabilidad neta. Esta distinción es fundamental, dado que un episodio con alto ingreso no necesariamente representa un episodio más rentable, pues para ello sería necesario incorporar información detallada de costos. Sin embargo, el estudio del ingreso por episodio constituye un primer paso para identificar grupos de alto peso económico, analizar la variabilidad dentro de diagnósticos prioritarios y orientar futuras revisiones tarifarias, contractuales u operativas.

Desde una perspectiva académica, esta práctica representó una experiencia significativa al permitir la aplicación de conocimientos matemáticos en un entorno real, interdisciplinario y de alta complejidad. El trabajo integró análisis de datos, estadística, aprendizaje automático, depuración de bases, interpretación clínica, modelamiento financiero y comunicación de resultados. Asimismo,

trascendió el alcance de una práctica convencional, al constituirse en un proyecto con potencial investigativo y posibilidades de continuidad en etapas posteriores de análisis predictivo, evaluación de costos y seguimiento institucional.

En síntesis, este informe presenta el desarrollo de una práctica empresarial orientada a evidenciar el papel que puede desempeñar un matemático en el análisis de información hospitalaria. La experiencia realizada en FOSCAL permitió mostrar que las matemáticas, al articularse con datos reales y necesidades institucionales concretas, pueden aportar herramientas valiosas para comprender fenómenos clínicos y financieros, apoyar la toma de decisiones y generar conocimiento aplicado. Así, el proyecto constituye no solo un ejercicio técnico de segmentación y análisis financiero, sino también una muestra del potencial profesional de la matemática en contextos distintos a los tradicionalmente asociados con la disciplina.

1. Generalidades

1.1 Planteamiento del problema

Los servicios de hospitalización atienden una población heterogénea en diagnósticos, trayectorias clínicas, complejidad, estancia y consumo de recursos. Esta diversidad se refleja en patrones diferenciados de ingreso por episodio y exige decisiones recurrentes sobre disponibilidad de camas, asignación de personal, rotación de unidades y sostenibilidad económica del servicio. En el caso de la Clínica FOSCAL, la información clínica, administrativa, operativa y financiera disponible para el periodo 2021–2024 ofrece una oportunidad para estudiar de manera estructurada esta heterogeneidad y su relación con el comportamiento financiero de los episodios hospitalarios.

Cuando no se cuenta con un proceso analítico reproducible que permita caracterizar perfiles homogéneos de episodios y relacionarlos sistemáticamente con variables financieras, la interpretación institucional puede apoyarse principalmente en análisis agregados, promedios históricos o revisiones puntuales. Aunque estos enfoques son útiles para el seguimiento general del servicio,

pueden no capturar diferencias relevantes entre subgrupos de pacientes con características clínicas y operativas distintas. En consecuencia, se dificulta la identificación de segmentos con alto peso económico, la caracterización de demandas operativas asociadas a la estancia y la comprensión de por qué episodios registrados con diagnósticos similares pueden presentar ingresos diferentes (1).

En este contexto, el problema no se limita a reconocer que existen diferencias entre diagnósticos, sino a comprender que un mismo diagnóstico principal puede agrupar episodios clínicamente distintos en edad, servicio tratante, duración de la estancia, procedimientos realizados y consumo de recursos. Por esta razón, analizar únicamente el diagnóstico de egreso puede resultar insuficiente para explicar la variabilidad observada en el ingreso por episodio. Desde una perspectiva metodológica, se requiere una estrategia que permita identificar perfiles clínicos más homogéneos y evaluar posteriormente si dichos perfiles presentan comportamientos financieros diferenciados. Esta aproximación no pretende establecer relaciones causales, sino aportar una lectura estructurada de la heterogeneidad hospitalaria y generar evidencia descriptiva y analítica que apoye la interpretación del desempeño financiero del servicio.

A partir de lo anterior, la pregunta orientadora del proyecto es: ¿Qué perfiles clínicos pueden identificarse mediante técnicas de agrupamiento y cómo se diferencian financieramente después de ajustar por estancia hospitalaria?

1.2 Justificación

La justificación de esta práctica empresarial se fundamenta en la necesidad de articular la formación matemática con problemas reales de gestión institucional en salud. En particular, el proyecto surge de un contexto hospitalario en el que la Clínica FOSCAL dispone de información clínica, administrativa, operativa y financiera asociada a los episodios de hospitalización, pero requiere metodologías analíticas que permitan organizar, interpretar y relacionar dicha información de manera reproducible. Por tanto, la práctica no se orientó únicamente a la aplicación de técnicas computacionales, sino a la construcción de un enfoque matemático y estadístico que permitiera apo-

yar la lectura institucional del servicio de hospitalización desde una perspectiva clínico-financiera.

1.2.1 Necesidad institucional

El servicio de hospitalización atiende una población heterogénea en diagnósticos, trayectorias clínicas, complejidad, duración de la estancia, servicios de atención y consumo de recursos. Esta heterogeneidad genera comportamientos diferenciados en los ingresos por episodio, los cuales pueden estar asociados a los procedimientos realizados, los materiales y medicamentos utilizados, los honorarios profesionales, la estancia hospitalaria y la ubicación asistencial del paciente dentro de la institución.

De acuerdo con el protocolo presentado al Comité de Ética de FOSCAL, la problemática institucional se relaciona con la ausencia de un proceso analítico reproducible que permita caracterizar perfiles homogéneos de pacientes y vincularlos de manera sistemática con su desempeño financiero. En dicho documento se plantea que, cuando esta caracterización no está disponible, la planeación tiende a apoyarse en promedios globales, reglas generales o análisis puntuales, lo cual limita la identificación de segmentos de pacientes con alto impacto económico y dificulta la comprensión de los factores clínicos y operativos que explican la variabilidad del ingreso por episodio (1).

En este contexto, el proyecto se justifica porque propone una metodología orientada a complementar la lectura tradicional del servicio de hospitalización mediante la segmentación clínica de episodios y el posterior análisis financiero de los grupos obtenidos. Esta aproximación permite pasar de una revisión general de la información a una lectura por perfiles clínico-operativos, sin afirmar que los indicadores institucionales existentes sean incorrectos, sino reconociendo que pueden fortalecerse mediante herramientas analíticas adicionales.

La necesidad institucional también se relaciona con el aprovechamiento de la información disponible. FOSCAL cuenta con registros anonimizados del periodo 2021–2024 que incluyen variables demográficas, clínicas, administrativas, operativas y financieras, tales como edad, sexo,

EPS, servicio, unidad médica, diagnóstico CIE-10, fechas de ingreso y egreso, estancia, ingreso por servicios, ingreso por materiales y medicamentos, ingreso por honorarios e ingreso total. Esta disponibilidad representa una oportunidad para construir una base analítica que permita estudiar la relación entre la casuística hospitalaria y el comportamiento financiero de los episodios de atención.

Por tanto, la práctica responde a una necesidad concreta: desarrollar un proceso reproducible de depuración, clasificación, segmentación y análisis financiero que permita identificar grupos de episodios hospitalarios con características clínicas similares y evaluar posteriormente cómo se diferencian en términos de ingreso y estancia. Esta información puede convertirse en un insumo útil para la gestión financiera, la priorización de análisis internos, la revisión de rutas asistenciales y la generación de evidencia institucional.

1.2.2 Pertinencia e impacto esperado

La pertinencia del proyecto se evidencia en tres niveles: institucional, académico y metodológico. A nivel institucional, el análisis permite aprovechar información rutinaria de hospitalización para construir una lectura más estructurada del comportamiento clínico y financiero de los episodios atendidos. Los resultados pueden apoyar la identificación de grupos con alta participación en el ingreso acumulado, grupos con mayor variabilidad financiera y diagnósticos que requieren revisión más detallada por su comportamiento económico.

A nivel académico, la práctica representa un ejercicio de aplicación directa de la formación matemática en un entorno real del sector salud. El proyecto permitió integrar herramientas de estadística, análisis de datos, aprendizaje no supervisado, modelamiento financiero y programación, demostrando que el matemático puede aportar no solo en escenarios docentes o teóricos, sino también en procesos institucionales de análisis, gestión y toma de decisiones. En este sentido, la experiencia fortalece la relación entre la Universidad Industrial de Santander y la Clínica FOSC-CAL, y abre la posibilidad de nuevas prácticas o proyectos interdisciplinarios desde la Escuela de Matemáticas.

A nivel metodológico, el proyecto aporta un flujo analítico reproducible que incluye control de calidad del dato, reclasificación clínica de diagnósticos, construcción de grupos clínicos, caracterización posterior por variables externas y análisis financiero ajustado por estancia. Esta estructura permite separar claramente las variables utilizadas para construir los grupos de aquellas empleadas para interpretarlos posteriormente, evitando que los grupos queden definidos artificialmente por mortalidad, ingreso, EPS, servicio o estancia.

El impacto esperado no consiste en reemplazar los procesos institucionales existentes, sino en aportar una herramienta complementaria para la comprensión del servicio de hospitalización. En particular, la segmentación clínica y el análisis financiero pueden ayudar a identificar perfiles de pacientes que concentran ingreso, presentan variabilidad relevante o requieren seguimiento por sus características clínicas y operativas. Estos resultados pueden ser utilizados como insumo para discusiones internas sobre planeación, auditoría, revisión tarifaria, análisis de rutas de atención y priorización de estudios posteriores de costos.

Además, el enfoque propuesto puede ser transferible a otras instituciones prestadoras de servicios de salud que cuenten con bases de datos estructuradas y registros hospitalarios comparables. Por esta razón, aunque el análisis se desarrolla en el contexto específico de FOSCAL, la metodología tiene potencial de adaptación a otros escenarios donde sea necesario relacionar información clínica, operativa y financiera de manera ordenada, trazable y reproducible.

1.2.3 Innovación y aporte del proyecto

El aporte principal del proyecto consiste en integrar, dentro de un mismo proceso analítico, la depuración de datos hospitalarios, la reconstrucción clínica de diagnósticos, la segmentación de episodios y el análisis financiero posterior de los grupos obtenidos. Esta integración permite estudiar la hospitalización no solo desde variables aisladas, sino desde perfiles clínico-financieros contruidos de manera sistemática.

En particular, el proyecto aporta tres elementos. Primero, una estrategia de segmentación

clínica basada en diagnósticos y variables demográficas, diseñada para identificar grupos interpretables de episodios hospitalarios. Segundo, una caracterización financiera de dichos grupos, orientada a evaluar diferencias en ingreso total, estancia y contribución acumulada. Tercero, una metodología reproducible que puede ser documentada, revisada y ajustada en futuras etapas del trabajo institucional.

Finalmente, esta práctica constituye un aporte formativo y profesional, porque muestra cómo la matemática puede contribuir a la solución de problemas reales en salud. La combinación de pensamiento abstracto, modelamiento, análisis estadístico y programación permitió construir una propuesta aplicada con utilidad institucional, fortaleciendo el papel del matemático como analista de datos en contextos clínicos y financieros.

2. Objetivos

2.1 Trazabilidad administrativa del proyecto

Trazabilidad administrativa del proyecto El desarrollo de esta práctica tuvo una trazabilidad académica e institucional compuesta por tres momentos: aprobación inicial de la propuesta por parte del comité de pregrado, solicitud de revisión ante el Comité de Ética de FOSCAL para el uso secundario de bases anonimizadas y ajuste del título del informe según el alcance finalmente desarrollado. Esta trazabilidad se menciona para dejar explícita la correspondencia entre el proyecto formulado, la autorización institucional y la versión final del informe. Como soporte de este proceso, se adjunta el documento de aprobación del proyecto por parte del Comité de Ética de FOSCAL en el Apéndice A. Los soportes documentales adicionales deben conservarse como respaldo administrativo del trabajo.

2.2 Objetivo general

Desarrollar e implementar un modelo analítico que combine técnicas de agrupamiento no supervisado y modelos explicativos ajustados para la segmentación de episodios hospitalarios según variables diagnósticas y demográficas, y para el análisis del ingreso financiero por episodio a partir de variables clínicas, operativas y financieras, con el fin de apoyar la gestión financiera y mejorar la interpretación operativa del servicio de hospitalización.

2.3 Objetivos específicos

1. Caracterizar la población de episodios hospitalarios mediante variables clínicas, operativas y financieras que describan el comportamiento de la hospitalización y del ingreso por episodio.
2. Analizar exploratoriamente el ingreso por episodio según variables demográficas, diagnósticas, administrativas y operativas, tales como edad, sexo, EPS, servicio, asignación de enfermería, unidad médica y estancia.
3. Segmentar episodios hospitalarios según variables diagnósticas y demográficas mediante técnicas de agrupamiento no supervisado, identificando grupos homogéneos de episodios con perfiles clínico-operativos diferenciados.
4. Caracterizar financieramente los segmentos identificados, relacionando los grupos clínicos con el ingreso por episodio, la estancia hospitalaria, la asignación de enfermería y la unidad médica.
5. Desarrollar y evaluar modelos explicativos ajustados del ingreso por episodio, basados en variables clínicas y operativas, para apoyar la gestión financiera y la interpretación operativa del servicio de hospitalización.

3. Marco referencial, estado del arte y consideraciones éticas

3.1 Marco teórico

El marco teórico se concentró en los conceptos directamente utilizados en el análisis: bases de datos estructuradas en salud, calidad del dato, agrupamiento no supervisado, reducción dimensional mediante *TruncatedSVD*, microagrupamiento con *MiniBatchKMeans*, agrupamiento jerárquico de Ward, métricas internas de evaluación y análisis financiero ajustado por estancia. Los

métodos alternativos de agrupamiento se mencionan únicamente como referencia metodológica, dado que no constituyeron la estrategia principal aplicada en los resultados.

Tabla 1

Síntesis del marco teórico y relación con el análisis aplicado

Concepto	Fundamento	Uso en el estudio
Bases de datos estructuradas en salud	El modelo relacional organiza la información en tablas, filas y columnas, lo que facilita integridad, normalización, consulta y reproducibilidad (2, 3).	Integración de variables demográficas, diagnósticas, administrativas, operativas y financieras de episodios hospitalarios.
Calidad del dato y preprocesamiento	Los registros secundarios pueden presentar problemas de completitud, consistencia, duplicidad, heterogeneidad textual y alta cardinalidad (1, 3).	Normalización de columnas, control de duplicados, faltantes, cardinalidad, recodificación Top- <i>N</i> , construcción de diccionarios y trazabilidad de variables derivadas.
Agrupamiento no supervisado	Busca organizar observaciones en grupos con semejanza interna y diferenciación externa, sin utilizar una variable respuesta supervisada (4, 5).	Segmentación clínica de episodios hospitalarios por similitud diagnóstica y demográfica, separando variables de construcción y variables de caracterización posterior.
Codificación <i>one-hot</i> y <i>TruncatedSVD</i>	La codificación <i>one-hot</i> representa variables categóricas como matriz binaria; <i>TruncatedSVD</i> reduce dimensionalidad en matrices dispersas (6-8).	Transformación de grupo etario, capítulo CIE-10, prefijo CIE-3 y sexo cuando aplicó, antes del agrupamiento.

Continúa en la siguiente página.

Concepto	Fundamento	Uso en el estudio
<i>MiniBatchKMeans</i> y Ward	<i>MiniBatchKMeans</i> resume grandes conjuntos mediante prototipos; Ward fusiona grupos minimizando el incremento de varianza interna (9-11).	Estrategia principal: microagrupamiento de episodios, cálculo de centroides y agrupamiento jerárquico sobre centroides en el espacio reducido.
Métricas internas y estabilidad	Silhouette, Calinski–Harabasz y Davies–Bouldin apoyan la evaluación de cohesión y separación; ARI y NMI evalúan concordancia entre particiones (12-16).	Selección multicriterio de k , validación con 30 corridas por cambio de semilla y 30 remuestreos del 80 %.
Análisis financiero ajustado	El ingreso hospitalario suele presentar asimetría y valores extremos; los modelos ajustados con errores robustos permiten comparar grupos controlando covariables (17-21).	Transformación logarítmica robusta del ingreso, ANCOVA con estancia continua, errores estándar HC3, comparaciones con Holm y residuos financieros ajustados.

3.2 Métodos alternativos considerados

Además de la estrategia principal, se revisaron métodos como Affinity Propagation, Mean Shift, Spectral Agrupamiento no supervisado, DBSCAN, HDBSCAN y OPTICS. Estos enfoques pueden ser útiles en otros escenarios porque permiten explorar ejemplares reales, modos de densidad, grafos de afinidad o agrupamientos con ruido. Sin embargo, no se adoptaron como metodología principal porque el objetivo aplicado requería grupos interpretables, de tamaño operativo, trazables y comparables en una matriz categórica dispersa. En consecuencia, el cuerpo del informe conserva la fundamentación detallada de los métodos efectivamente usados: codificación *one-hot*,

TruncatedSVD, *MiniBatchKMeans*, *Ward*, métricas internas, estabilidad por remuestreo y modelos financieros ajustados.

3.3 Componente financiero: del episodio clínico a la caracterización económica

En las instituciones prestadoras de servicios de salud, el desempeño económico se relaciona con la complejidad asistencial, la intensidad de uso de recursos, la duración de la estancia y la estructura de costos e ingresos por evento. La estancia hospitalaria es un determinante operativo relevante, pues se asocia con variaciones en costos, eficiencia y ocupación (22). Desde un enfoque de gestión basada en valor, se promueve analizar resultados clínicos en relación con el costo o ingreso por episodio, fortaleciendo metodologías que aproximen la realidad económica a nivel de paciente (23, 24).

El protocolo del proyecto incorpora variables financieras de ingreso por episodio, tales como `ingreso_med_x_epi`, `ingreso_mat_x_epi`, `ingreso_hon_x_epi`, `ingreso_serv_x_epi` e `ingreso_total` (1). En este informe estas variables se utilizan para caracterizar grupos, identificar segmentos con alta concentración de ingreso y priorizar análisis de variabilidad financiera. El estudio no estima utilidad ni rentabilidad neta porque no integra costos completos.

3.4 Estado del arte

El aprendizaje no supervisado en salud ha crecido por la disponibilidad de registros clínicos y administrativos. La literatura muestra aplicaciones en segmentación poblacional, estratificación de riesgo, diseño de intervenciones diferenciadas y planeación de servicios (25, 26). Una revisión sistemática sobre segmentación poblacional basada en datos resalta que estos métodos permiten dividir poblaciones heterogéneas en grupos más homogéneos, siempre que se evalúe no solo la validez estadística sino también la interpretabilidad y la accionabilidad clínica (27).

Las revisiones metodológicas insisten en que el desempeño del agrupamiento depende de la

calidad del dato, la selección de variables, el manejo de variables mixtas y la validación mediante criterios cuantitativos y juicio experto (26). En este trabajo, la segmentación no se interpretó como clasificación diagnóstica definitiva, sino como una herramienta analítica para organizar episodios hospitalarios y relacionarlos posteriormente con variables clínicas, operativas y financieras.

3.5 Consideraciones éticas, privacidad y uso secundario de datos

El proyecto utilizó exclusivamente bases de datos anonimizadas provenientes de los sistemas de información institucionales de FOSCAL. No hubo contacto con personas, intervención sobre pacientes ni modificación de conductas o tratamientos clínicos. Los resultados se reportan de forma agregada, sin posibilidad de identificación individual.

El estudio se acogió a principios internacionales de investigación en salud, incluyendo la Declaración de Helsinki, el Informe Belmont y las Pautas Éticas Internacionales del CIOMS y la OMS. Asimismo, se acoge a la normativa colombiana sobre protección de datos personales y manejo de historias clínicas, en particular la Ley 1581 de 2012, el Decreto 1377 de 2013, la Ley 1266 de 2008 y la Resolución 1995 de 1999 (28-30).

La investigación se planteó como un estudio sin riesgo directo, dado que emplea bases anonimizadas y no involucra intervención ni interacción con personas. El análisis se fundamenta en los principios de autonomía, beneficencia, no maleficencia y justicia: no identifica individuos, busca generar beneficio institucional y social, minimiza riesgos de reidentificación mediante controles de

anonimización y seguridad, y analiza los registros de manera agregada y equitativa.

4. Metodología

4.1 Diseño, contexto empresarial y alcance de la práctica

La práctica empresarial se desarrolló en FOSCAL a partir de una necesidad institucional: transformar registros rutinarios de hospitalización, generados en los procesos clínicos, operativos y financieros, en insumos analíticos que permitieran interpretar perfiles de episodios hospitalarios y su comportamiento financiero. La realidad institucional se caracteriza por heterogeneidad diagnóstica, diversidad de servicios y unidades médicas, variabilidad en la estancia, multiplicidad de aseguradores y diferencias relevantes en el ingreso por episodio. En este contexto, la intervención realizada fue de tipo analítico-aplicado: no consistió en modificar procesos asistenciales, sino en construir una ruta reproducible de depuración, agrupamiento, caracterización y análisis financiero.

La unidad disponible para el análisis fue el episodio hospitalario. No se contó con un identificador longitudinal de paciente que permitiera evaluar dependencia entre episodios de un mismo individuo, reingresos independientes o trayectorias longitudinales. Por tanto, las frecuencias, proporciones, mortalidad, estancia e ingresos se interpretan a nivel de episodio y no como estimaciones individuales por paciente.

La práctica se organizó en dos componentes articulados. Primero, se construyeron grupos clínico-diagnósticos mediante agrupamiento no supervisado. Segundo, se caracterizó financieramente la solución obtenida mediante ingreso total, estancia, contribución acumulada, medias ajustadas y residuos financieros. La separación entre fase clínica y fase financiera evitó circularidad metodológica: las variables usadas para interpretar los grupos no fueron utilizadas por el algoritmo para construirlos.

4.2 Datos, cohorte de análisis y flujo metodológico

La información correspondió a registros secundarios anonimizados del servicio de hospitalización de FOSCAL para el periodo 2021–2024. La extracción fue realizada por la Unidad de Inteligencia Empresarial de FOSCAL y entregada al equipo del proyecto sin identificación individual. El uso de datos secundarios exige documentar fuente, unidad de análisis, criterios de inclusión y transformaciones aplicadas (31, 32).

El periodo 2021–2024 se utilizó para análisis exploratorio general, auditoría temporal de calidad del dato, sensibilidad de recodificaciones y análisis bivariado preliminar. Los análisis principales de agrupamiento clínico y caracterización financiera se restringieron a la cohorte 2024 sin COVID-19. Esta decisión respondió a un criterio operativo definido con acompañamiento médico: trabajar con el año completo más reciente disponible y reducir la influencia de patrones asistenciales asociados a la pandemia. La decisión no se presenta como prueba de representatividad interanual ni como demostración de calidad superior del dato en 2024.

Tabla 2*Flujo metodológico general de datos*

Etapas	Descripción metodológica
Base inicial 2021–2024	Consolidación de registros hospitalarios anonimizados provenientes del ERP institucional SAP, con información clínica, administrativa, operativa y financiera.
Auditoría temporal	Evaluación de indicadores específicos de calidad diagnóstica, completitud, cardinalidad y consistencia de variables.
Selección de cohorte 2024	Priorización del año completo más reciente por pertinencia institucional y recomendación clínica-operativa.
Exclusión COVID-19	Separación de episodios COVID-19 2024 y sensibilidad descriptiva sobre tamaño, mortalidad, estancia y composición relativa.
Base clínica final	Cohorte 2024 sin COVID-19, utilizada para reclasificación clínica, construcción de macrogrupos y agrupamiento no supervisado.
Base financiera	Integración posterior de ingreso total y estancia para ANCOVA, contribución acumulada, medias ajustadas y residuos financieros.

4.3 Preparación de variables, recodificación y análisis bivariado preliminar

La base contenía variables clínicas, demográficas, administrativas, operativas y financieras: identificador de episodio, fechas de ingreso y egreso, clase de movimiento al alta, asignación de enfermería, servicio, unidad médica, estancia, estancia facturada, edad, grupo de edad, sexo, EPS, departamento y ciudad de residencia, código CIE-10, nombre diagnóstico, gran causa de morbilidad, fecha de muerte e ingresos por componentes financieros. Las variables de mortalidad, estancia, ingreso total, EPS, servicio, asignación de enfermería y unidad médica se reservaron para caracterización posterior o análisis financiero, no para construir grupos clínicos.

Tabla 3*Variables disponibles/preparadas y uso metodológico*

Variable	Tipo	Fuente	Uso en el estudio
episodio	Identificador	ERP institucional SAP / base de hospitalización	Unidad de análisis y control de duplicados.
edad y grupo de edad	Cuantitativa / categórica	Base institucional y variable derivada	Exploración, agrupamiento mediante grupo etario y caracterización posterior.
sexo	Categórica nominal	Base institucional	Agrupamiento en infecciosa no COVID y caracterización posterior.
diagnóstico CIE-10 y nombre diagnóstico	Categórica codificada / textual	Base institucional	Reclasificación clínica, control de calidad, derivación de capítulo CIE-10 y CIE-3.
capítulo CIE-10 y CIE-3	Categóricas derivadas	Derivadas del código CIE-10	Variables principales para agrupamiento clínico.
gran causa de morbilidad	Categórica nominal	Base institucional	Control de calidad y comparación con clasificación reconstruida.
estancia	Cuantitativa	Base institucional	Covariable continua del análisis financiero y caracterización descriptiva.
ingreso total	Cuantitativa	Base financiera institucional	Respuesta financiera principal; no se usó para formar grupos.

Continúa en la siguiente página.

Variable	Tipo	Fuente	Uso en el estudio
EPS, servicio, asignación de enfermería y unidad médica	Categóricas	Base institucional	Caracterización posterior, análisis bivariado y lectura institucional.
mortalidad	Binaria derivada	Derivada de fecha de muerte	Desenlace externo post hoc; no se usó para formar grupos.

La preparación inicial incluyó normalización de nombres de columnas, resolución de columnas clave mediante listas de candidatos, exportación de diccionarios, verificación de duplicados exactos, control de duplicados por episodio, cálculo de faltantes, cardinalidad y análisis exploratorio de edad, estancia e ingreso. Estos procedimientos permitieron identificar fragmentación categórica, valores extremos y necesidad de transformaciones financieras (33).

Las variables categóricas de alta cardinalidad se recodificaron mediante reglas Top- N , conservando las categorías más frecuentes y agrupando el resto en una categoría residual. Para una variable categórica X , la recodificación fue:

$$x_i^{(N)} = \begin{cases} x_i, & \text{si } x_i \text{ pertenece a las } N \text{ categorías más frecuentes,} \\ \text{categoría residual,} & \text{en caso contrario.} \end{cases} \quad (4.1)$$

Esta regla se aplicó con trazabilidad mediante diccionarios de categoría original, categoría recodificada, frecuencia, porcentaje, porcentaje acumulado y regla aplicada. Para diagnóstico de egreso se evaluaron versiones Top 30, Top 50 y Top 100; para asignación de enfermería, UO médica, unidad médica, EPS y ciudad se evaluaron versiones Top 12. Los resultados empíricos de cobertura y sensibilidad se reportan en el Capítulo 5 y en apéndices.

El análisis bivariado preliminar se realizó sobre la base 2021–2024 con finalidad exploratoria y metodológica. Para cada par de variables cualitativas se calculó prueba chi-cuadrado, V de Cramér,

grados de libertad, porcentaje de frecuencias esperadas menores a 5, mínimo esperado, corrección Benjamini–Hochberg e intervalos bootstrap para V cuando aplicó (34-38). El tamaño efectivo por cruce fue:

$$n_{XY} = \#\{i : X_i \neq \text{faltante}, Y_i \neq \text{faltante}\}. \quad (4.2)$$

El criterio $V \geq 0,10$ se usó como regla operativa de priorización, no como umbral universal de relevancia. Los cruces tautológicos o jerárquicamente dependientes se clasificaron como controles técnicos y no como hallazgos sustantivos.

4.4 Reclasificación clínica y agrupamiento no supervisado

Antes del agrupamiento, se reconstruyó la clasificación clínica de los episodios a partir del código CIE-10 y del nombre diagnóstico. La reclasificación siguió reglas jerárquicas para asignar una única categoría: COVID-19, materna-obstétrica, perinatal-neonatal, lesiones, infecciosa no COVID, mal definidas o inespecíficas, o no transmisible. Esta etapa fue una reclasificación operativa reproducible, no una validación clínica individual ni una auditoría de historia clínica (39).

La variable de discrepancia clínica se construyó después de la reclasificación para identificar diferencias entre clasificación original y reconstruida. Su interpretación se restringió a consistencia entre esquemas de clasificación; no se interpretó como error diagnóstico individual.

La cohorte 2024 sin COVID-19 se estratificó por macrogrupo clínico. Solo los macrogrupos No transmisible e Infecciosa no covid fueron sometidos a agrupamiento algorítmico por su tamaño y heterogeneidad interna. Los macrogrupos materna-obstétrica, perinatal-neonatal, lesiones y mal definidas o inespecíficas se conservaron como grupos clínicos únicos.

Las variables cualitativas se transformaron mediante codificación *one-hot*. Posteriormente se aplicó *TruncatedSVD* para reducir la matriz dispersa a una representación continua de menor dimensión (6, 7). La configuración principal utilizó 10 componentes SVD, documentada como decisión parsimoniosa de compromiso entre factibilidad computacional, conservación parcial de estructura categórica e interpretabilidad posterior.

La estrategia de agrupamiento combinó *MiniBatchKMeans* y Ward. Primero se formaron microgrupos con:

$$n_{\text{micro}} = \min\{120, \max(20, \lfloor n/30 \rfloor)\}, \quad (4.3)$$

donde n corresponde al número de episodios del macrogrupo. Luego se calcularon centroides en el espacio SVD y sobre ellos se aplicó agrupamiento jerárquico aglomerativo con enlace de Ward y distancia euclidiana (9-11). Cada episodio heredó el grupo final del microgrupo al que pertenecía.

La selección de k se basó en un criterio multicriterio: métricas internas, tamaño mínimo de grupo, estabilidad, comparabilidad entre macrogrupos e interpretabilidad clínica. Se calcularon Silhouette, Calinski–Harabasz, Davies–Bouldin y tamaño mínimo para $k = 2, \dots, 8$ (12-14). La validación final de estabilidad se realizó con 30 corridas por cambio de semilla y 30 remuestreos del 80 %, evaluadas mediante ARI y NMI (15, 16).

4.5 Caracterización clínica, análisis financiero y residuos ajustados

Una vez contruidos los grupos, se realizó caracterización externa/post hoc con variables que no participaron en la formación de la partición: edad, sexo, servicio, asignación de enfermería, unidad médica, diagnósticos principales, mortalidad, estancia e ingreso. La mortalidad intragrupo se calculó como:

$$\text{Mortalidad del grupo} = \frac{\text{Fallecimientos en el grupo}}{\text{Total de episodios del grupo}} \times 100, \quad (4.4)$$

y el aporte al total de fallecimientos como:

$$\text{Aporte a fallecimientos} = \frac{\text{Fallecimientos en el grupo}}{\text{Total de fallecimientos de la cohorte}} \times 100. \quad (4.5)$$

El análisis financiero utilizó como respuesta principal el ingreso total del episodio. No se interpretó como costo, utilidad, margen ni rentabilidad. Antes del modelo se compararon trans-

formaciones candidatas: escala original, logaritmo, logaritmo robusto, Box–Cox, Yeo–Johnson y winsorización con logaritmo robusto (17, 18). La respuesta principal fue:

$$Y_i^* = \frac{\log(Y_i + c) - \text{Mediana}[\log(Y + c)]}{\text{IQR}[\log(Y + c)]}, \quad (4.6)$$

donde c garantizó positividad cuando fuera necesario. La estancia se incorporó como covariable continua estandarizada:

$$\text{estancia}_{z,i} = \frac{\text{estancia}_i - \bar{\text{estancia}}}{s_{\text{estancia}}}. \quad (4.7)$$

Para comparar grupos se ajustó un ANCOVA con respuesta transformada y estancia continua. Se evaluó un modelo aditivo y un modelo con interacción grupo clínico por estancia:

$$Y_i^* = \beta_0 + \alpha_{\text{grupo}_i} + \beta_1 \text{estancia}_{z,i} + \gamma_{\text{grupo}_i} \text{estancia}_{z,i} + \varepsilon_i. \quad (4.8)$$

Dado el desbalance entre grupos, la asimetría financiera y la heterocedasticidad esperable, la inferencia se apoyó en errores estándar robustos HC3 y corrección de Holm para comparaciones posteriores (19-21, 40). Los resultados se interpretaron como diferencias ajustadas en escala transformada y como medias ajustadas o estimaciones marginales ajustadas, no como predicciones supervisadas ni efectos causales.

Dentro del grupo de mayor peso financiero se calculó un residuo financiero ajustado por diagnóstico y estancia:

$$e_i = Y_{\text{observado},i}^* - \widehat{Y}_{\text{esperado},i}^*. \quad (4.9)$$

Un residuo positivo indicó ingreso observado superior al esperado en la escala transformada, dada la combinación de diagnóstico y estancia incluida en el modelo. Esta medida se utilizó como criterio exploratorio de priorización; no se interpretó como cobro indebido, sobreutilización, utilidad ni rentabilidad.

4.6 Reproducibilidad computacional y trazabilidad del código

El procesamiento se realizó en Python porque el flujo requería integración de bases, manejo de matrices dispersas, codificación *one-hot*, reducción dimensional, agrupamiento escalable, modelos estadísticos y exportación sistemática de tablas y figuras. Python permitió articular `pandas`, `numpy`, `scipy`, `scikit-learn`, `statsmodels`, `matplotlib` y `seaborn` en cuadernos reproducibles (8, 41-46). R habría sido una alternativa válida para varios procedimientos estadísticos; sin embargo, la elección de Python respondió a continuidad del flujo computacional, compatibilidad con matrices dispersas y uso de Google Colab/Jupyter para trazabilidad del procesamiento.

La reproducibilidad desde Google Colab se garantizó mediante: fijación de semillas, registro de versiones de librerías, exportación de tablas intermedias, diccionarios de recodificación, figuras diagnósticas, métricas de validación y separación de notebooks por etapa analítica. El flujo computacional general fue:

Tabla 4*Pseudocódigo general del flujo computacional utilizado*

Paso	Procedimiento
1	Cargar base anonimizada 2021–2024 y normalizar nombres de columnas.
2	Verificar duplicados, faltantes, cardinalidad y consistencia básica de variables.
3	Construir variables derivadas: fallecimiento, capítulo CIE-10, CIE-3, grupos etarios, variables normalizadas y recodificaciones Top- N .
4	Ejecutar análisis exploratorio y bivariado preliminar con chi-cuadrado, V de Cramér, corrección Benjamini–Hochberg y bootstrap.
5	Reconstruir clasificación clínica mediante reglas jerárquicas CIE-10/texto y excluir COVID-19 para la cohorte principal 2024.
6	Codificar variables categóricas con <i>one-hot</i> , aplicar <i>TruncatedSVD</i> , formar microgrupos con <i>MiniBatchKMeans</i> y agrupar centroides con Ward.
7	Evaluar k , estabilidad con 30 semillas y 30 remuestreos, y caracterizar grupos con variables externas/post hoc.
8	Transformar ingreso, ajustar ANCOVA con estancia continua, calcular medias ajustadas, contribución acumulada y residuos financieros.
9	Exportar tablas, figuras, diccionarios, métricas y bases derivadas para revisión y reproducción.

Tabla 5*Resumen de reproducibilidad computacional del análisis*

Componente	Detalle
Lenguaje de programación	Python 3.12.13.
Sistema operativo	Linux 6.6.122, arquitectura x86_64, glibc 2.35.
Ambiente de ejecución	Google Colab / Jupyter, según entorno de ejecución del notebook.
Librerías principales	pandas 2.2.2; numpy 2.0.2; scipy 1.16.3; scikit-learn 1.6.1; statsmodels 0.14.6; matplotlib 3.10.0; seaborn 0.13.2.
Semillas	random_state = 42 en agrupamiento principal; RANDOM_SEED = 123 en procedimientos financieros con componente aleatorio.
Validación de estabilidad	30 corridas por cambio de semilla y 30 remuestreos del 80 % para macro-grupos subdivididos.
Modelo financiero	ANCOVA con interacción grupo clínico por estancia; errores estándar HC3; corrección Holm; residuos ajustados por diagnóstico y estancia.
Archivos de salida	Tablas CSV/XLSX, figuras PNG, diccionarios de recodificación, métricas de validación y bases analíticas derivadas.

5. Preparación de datos y análisis preliminar

5.1 Estructura general, variables analíticas e integridad de la base

Se cargó la base consolidada de morbilidad y mortalidad correspondiente a hospitalizaciones ocurridas durante el periodo 2021–2024. La base original estuvo conformada por 94.134 registros y 32 variables suministradas por la institución. Durante la preparación analítica se generaron variables derivadas para estandarización textual, recodificación, análisis bivariado, reconstrucción clínica y agrupamiento. Por tanto, en este capítulo se distingue entre las variables originales de la base

institucional y las variables analíticas construidas durante el procesamiento.

Para el análisis exploratorio y bivariado preliminar se definieron variables objetivo que cubren dimensiones demográficas, asistenciales, administrativas, geográficas, clínicas, diagnósticas, de desenlace y financieras. Estas variables permitieron documentar la estructura de la base antes de restringir los análisis principales de segmentación clínica a la cohorte 2024 sin COVID-19.

Tabla 6

Dimensiones analíticas y variables objetivo iniciales

Dimensión	Variables consideradas
Demográfica	Edad, grupo de edad y sexo.
Asistencial y operativa	Servicio, asignación de enfermería, unidad médica y UO médica.
Administrativa	EPS.
Geográfica	Departamento y ciudad.
Clínica y diagnóstica	Código CIE-10, nombre del diagnóstico de egreso, capítulo CIE-10, prefijo CIE-3, gran causa de morbilidad y clasificación clínica reconstruida.
Desenlace	Fecha de muerte y variable binaria de fallecimiento.
Financiera	Ingreso total y estancia hospitalaria para la evaluación preliminar de transformaciones financieras.

La verificación estructural confirmó la presencia de las variables necesarias para el análisis planificado. Algunas variables fueron utilizadas exclusivamente para caracterización posterior o control metodológico, y no para la formación de grupos clínicos. En particular, mortalidad, estancia, ingreso total, EPS, servicio, unidad médica y UO médica se reservaron para análisis descriptivos, bivariados o financieros posteriores. Esta decisión evitó que los grupos clínicos fueran construidos directamente a partir de variables de desenlace, administrativas u operativas.

Se realizó una verificación de duplicados exactos a nivel de fila. El procedimiento no identificó registros repetidos, lo cual respaldó la integridad básica de la información y permitió continuar con las etapas de análisis exploratorio y recodificación. En cuanto a los tipos de datos, la mayoría de variables objetivo se encontraba almacenada como texto categórico; la edad estaba

registrada como variable numérica y la fecha de muerte como variable temporal.

También se efectuó una auditoría inicial sobre las variables objetivo considerando completitud, cardinalidad y presencia de categorías poco frecuentes. La mayoría de las variables analíticas presentó niveles altos de completitud. En contraste, causa básica de muerte y fecha de muerte registraron valores faltantes superiores al 97 %, situación esperable por su aplicabilidad restringida al subgrupo de episodios con fallecimiento registrado. Por esta razón, estas variables no se interpretaron como variables generales de caracterización de toda la cohorte.

Asimismo, se identificaron variables con alta cardinalidad, especialmente el diagnóstico de egreso, la ciudad, el prefijo CIE-3 y la fecha de muerte. Este grado de fragmentación justificó una etapa posterior de recodificación documentada y la construcción de versiones Top- N , con el objetivo de obtener tablas de contingencia más legibles y reducir la dispersión de categorías de baja frecuencia. Las visualizaciones descriptivas de faltantes y cardinalidad se presentan en los Complementos del Capítulo 5, Figuras 7 y 8.

5.2 Exploración preliminar, concentración diagnóstica y recodificación categórica

Antes de cualquier recodificación, se realizó un análisis exploratorio de variables cuantitativas y cualitativas. En las variables cuantitativas se revisaron distribución, asimetría, presencia de valores extremos y comportamiento general. En las variables cualitativas se evaluó concentración de categorías, categorías raras y pertinencia de construir versiones recodificadas. Las distribuciones categóricas completas fueron exportadas desde el notebook como figuras y tablas de frecuencia. Las visualizaciones de carácter descriptivo se trasladaron a los Complementos del Capítulo 5, Figuras 9, 10 y 11.

La concentración diagnóstica mostró una estructura de alta cardinalidad: un número reducido de diagnósticos agrupó una proporción importante de episodios, mientras que muchas categorías presentaron baja frecuencia. Este comportamiento sustentó la construcción de variables diagnósticas Top- N y la evaluación de sensibilidad frente a distintos puntos de corte.

Con el propósito de mejorar la estabilidad descriptiva, la legibilidad de las tablas de contingencia y la trazabilidad del análisis, se implementó una etapa de recodificación y estandarización de variables categóricas. Esta etapa incluyó normalización textual, construcción de variables derivadas y generación de versiones Top- N por frecuencia.

En primer lugar, se construyó la variable dicotómica de fallecimiento a partir de la existencia de fecha de muerte registrada. En la base 2021–2024 se identificaron 2.088 episodios con fallecimiento, equivalentes al 2,22 %, y 92.046 episodios sin fallecimiento registrado, equivalentes al 97,78 %.

La categoría NO_APLICA en la variable derivada de fecha de muerte no se interpretó como una categoría temporal sustantiva, sino como una codificación operativa para identificar episodios sin fallecimiento registrado. Por tanto, esta categoría no fue utilizada para analizar distribuciones temporales de mortalidad; para ese propósito, los episodios sin evento deben excluirse o analizarse mediante la variable binaria de fallecimiento.

Tabla 7
Recodificación de variables categóricas por frecuencia acumulada

Variable original	Variable recodificada	Criterio	Cat. orig.	Cat. finales	% conservado	% residual
dx_nombre_norm	dx_egreso_top30	Top 30	3785	31	39,2600	60,7400
dx_nombre_norm	dx_egreso_top50	Top 50	3785	51	47,5206	52,4794
dx_nombre_norm	dx_egreso_top100	Top 100	3785	101	59,5948	40,4052
asign_enfermeria_norm	asign_enfermeria_norm_top12	Top 12	21	13	90,2416	9,7584
uo_medico_norm	uo_medico_norm_top12	Top 12	80	13	93,1714	6,8286
unidad_medica_norm	unidad_medica_norm_top12	Top 12	80	13	93,1714	6,8286
eps_norm	eps_norm_top12	Top 12	80	13	96,1066	3,8934
ciudad_norm	ciudad_norm_top12	Top 12	302	13	89,1878	10,8122

La Tabla 7 resume las reglas de recodificación por frecuencia acumulada utilizadas para reducir la cardinalidad de variables categóricas. En todos los casos, las categorías no incluidas dentro del respectivo punto de corte Top- N fueron agrupadas en una categoría residual. Esta

decisión permitió conservar las categorías más frecuentes, mejorar la legibilidad de las tablas de contingencia y reducir la dispersión generada por categorías de baja frecuencia.

La tabla muestra que las variables operativas y administrativas lograron coberturas acumuladas superiores al 89 % con las configuraciones Top 12. En contraste, el diagnóstico de egreso presentó mayor dispersión: incluso con Top 50, el 52,48 % de episodios quedó agrupado en la categoría residual. Este resultado no invalida la recodificación diagnóstica, pero exige interpretar dicha categoría como una agrupación heterogénea de diagnósticos poco frecuentes. La distribución gráfica de las categorías principales del diagnóstico de egreso después de la recodificación Top 50 se presenta en los Complementos del Capítulo 5, Figura 13.

Los diccionarios completos de recodificación se presentan como apéndices metodológicos. Dichos diccionarios incluyen categoría original, categoría recodificada, frecuencia absoluta, porcentaje, porcentaje acumulado y regla aplicada, permitiendo reconstruir las variables derivadas a partir de la base original.

5.3 Definición de la cohorte 2024 sin COVID-19

La cohorte 2024 incluyó inicialmente 26.280 episodios. Después de aplicar la clasificación clínica reconstruida, se identificaron 14 episodios correspondientes a COVID-19. La exclusión de estos episodios se evaluó mediante sensibilidad, comparando la cohorte con y sin COVID-19 en términos de tamaño, mortalidad y estancia hospitalaria.

Tabla 8
Comparación de cohortes de episodios hospitalarios 2024 según exclusión COVID-19

Escenario	<i>n</i>	% cohorte	Fallec.	Mortalidad %	Estancia mediana	Lectura
Cohorte 2024 con COVID	26.280	100,0000	469	1,7846	2,0	Referencia.
Cohorte 2024 sin COVID	26.266	99,9467	469	1,7856	2,0	Cambio mínimo.
Episodios COVID excluidos	14	0,0533	0	0,0000	2,0	Tamaño marginal.

La exclusión de los 14 episodios COVID-19 no modificó el número absoluto de fallecimientos y produjo cambios mínimos en la mortalidad porcentual y en la estancia mediana. Adicionalmente, se revisó la composición relativa de los macrogrupos clínicos con y sin COVID-19. Las diferencias porcentuales fueron marginales en los macrogrupos no COVID, lo que indica que la exclusión operó principalmente como un ajuste de delimitación de cohorte y no como una modificación sustantiva de la composición general.

Tabla 9

Sensibilidad de la exclusión COVID-19 sobre la composición relativa de macrogrupos

Macrogrupo	% con COVID	% sin COVID	Diferencia pp. sin-con
NO_TRANSMISIBLE	44,2846	44,3082	0,0236
INFECCIOSA_NO_COVID	20,2816	20,2924	0,0108
MATERNA_OBSTETRICA	16,0921	16,1007	0,0086
MAL_DEFINIDAS_INESPECIFICAS	12,6636	12,6704	0,0068
LESIONES	4,3113	4,3136	0,0023
PERINATAL_NEONATAL	2,3135	2,3148	0,0012
COVID-19	0,0533	0,0000	-0,0533

La visualización complementaria de esta sensibilidad se trasladó a los Complementos del Capítulo 5, Figura 12. Con base en estos resultados, la cohorte final para agrupamiento clínico quedó conformada por 26.266 episodios hospitalarios de 2024 sin COVID-19.

5.4 Sensibilidad de recodificaciones y análisis bivariado global

La elección de la configuración principal de recodificación no se basó únicamente en conveniencia. Se realizó un análisis de sensibilidad comparando varias configuraciones Top-*N*, especialmente para diagnóstico de egreso, asignación de enfermería y UO médica. La sensibilidad evaluó cobertura acumulada, valor promedio de *V* de Cramér, asociación principal y correlación

del ranking frente a la configuración final.

Tabla 10

Análisis de sensibilidad y configuración de variables categóricas

Configuración	Variable	Cobertura %	V promedio	Corr. ranking	Conclusión
Top 30	Diagnóstico de egreso	39,2600	0,4712	0,9833	Sensibilidad
Top 50	Diagnóstico de egreso	47,5206	0,4879	1,0000	Configuración final
Top 100	Diagnóstico de egreso	59,5948	0,5086	0,8909	Sensibilidad
Top 10	UO médica	90,9799	0,4942	1,0000	Sensibilidad
Top 12	UO médica	93,1714	0,4879	1,0000	Configuración final
Top 15	UO médica	95,3757	0,4806	0,9879	Sensibilidad
Top 10	Asignación enfermería	84,3946	0,4886	0,9879	Sensibilidad
Top 12	Asignación enfermería	90,2416	0,4879	1,0000	Configuración final
Top 15	Asignación enfermería	96,6601	0,5019	0,9273	Sensibilidad

Los resultados mostraron que las configuraciones Top 50 para diagnóstico y Top 12 para UO médica y asignación de enfermería ofrecieron un equilibrio adecuado entre cobertura, estabilidad del ranking de asociaciones y legibilidad de las tablas de contingencia. Estos puntos de corte se asumieron como criterios operativos de reducción de cardinalidad, no como umbrales universales.

El análisis bivariado preliminar se realizó sobre la base consolidada 2021–2024, con 94.134 registros, según lo descrito en la Sección 4.3. Su propósito fue exploratorio y no intervino directamente en la construcción de los grupos clínicos finales. Se evaluaron 91 pares de variables categóricas mediante chi-cuadrado, V de Cramér, ajuste Benjamini–Hochberg y revisión de condiciones asintóticas.

Debido al tamaño muestral, la significancia estadística se interpretó con cautela. La priorización de asociaciones se basó principalmente en el tamaño de efecto, la validez de las condiciones esperadas y la plausibilidad clínica u operativa, diferenciando así entre significancia estadística y relevancia práctica.

Tabla 12*Asociaciones con mayor V de Cramér e intervalos bootstrap disponibles*

Variable 1	Variable 2	Dim.	V	IC 95 % inf.	IC 95 % sup.	Asintótico aceptable
gran_causa_norm	cie_capitulo	7x26	0,8834	0,8815	0,8844	No
servicio_norm	uo_medico_norm_top12	5x13	0,6294	0,6227	0,6313	Sí
servicio_norm	unidad_medica_norm_top12	5x13	0,6294	0,6227	0,6313	Sí
gran_causa_norm	dx_egreso_top50	7x51	0,6204	0,6136	0,6249	No
grupo_edad_norm	servicio_norm	17x5	0,5952	0,5904	0,5970	Sí
servicio_norm	cie_capitulo	5x26	0,5489	0,5454	0,5540	No
servicio_norm	asignacion_enfermeria_norm_top12	5x13	0,5308	0,5238	0,5403	Sí
dx_egreso_top50	cie_capitulo	51x26	0,5129	0,5078	0,5169	No
servicio_norm	dx_egreso_top50	5x51	0,4945	0,4912	0,5027	No

Nota. Todos los cruces se calcularon sobre $n = 94,134$ registros. La columna “Dim.” indica la dimensión de la tabla de contingencia. Se excluyó de esta tabla el cruce entre `uo_medico_norm_top12` y `unidad_medica_norm_top12`, dado que presentó $V = 1$, lo que sugiere equivalencia o duplicación operativa entre ambas variables y no una asociación sustantiva independiente.

Los intervalos bootstrap disponibles muestran estimaciones muy precisas para los principales tamaños de efecto, lo cual es esperable dada la magnitud de la muestra. No obstante, la precisión estadística no elimina la necesidad de revisar dependencia conceptual, condiciones asintóticas y pertinencia sustantiva de cada cruce.

Tabla 13*Cruces no interpretados como hallazgos sustantivos por dependencia conceptual*

Variable 1	Variable 2	V	Dim.	Tratamiento
uo_medico_norm_top12	unidad_medica_norm_top12	1,0000	13x13	No se interpreta como hallazgo sustantivo; se clasifica como dependencia conceptual o duplicación operativa.
dx_egreso_top50	cie_capitulo	0,5129	51x26	No se interpreta como asociación independiente, porque el capítulo CIE-10 deriva jerárquicamente del diagnóstico.

A partir del análisis bivariado se seleccionaron cuatro cruces para profundización post hoc. La selección no se basó únicamente en el valor de V , sino en la combinación entre magnitud del efecto, relevancia clínica u operativa, utilidad para interpretar la base y pertinencia institucional. Los cruces seleccionados fueron: servicio con diagnóstico de egreso Top 50, servicio con fallecimiento, servicio con asignación de enfermería Top 12 y asignación de enfermería Top 12 con EPS Top 12. La descomposición celular de estos cruces se presenta en el Capítulo 6.

5.5 Preparación preliminar de variables para el agrupamiento clínico

Para preparar el agrupamiento clínico definitivo se construyeron variables diagnósticas derivadas del código CIE-10, principalmente capítulo CIE-10 y prefijo CIE-3. También se consideró grupo de edad y, en el macrogrupo infecciosa no COVID, sexo. Esta sección no presenta la solución final de agrupamiento, sino las verificaciones preliminares que justifican su factibilidad: distribución diagnóstica, sensibilidad de componentes SVD y evaluación inicial de valores de k . Las

visualizaciones preliminares de grupo de edad y prefijo CIE-3 se trasladaron a los Complementos del Capítulo 5, Figuras 14 y 15.

Tabla 14

Distribución de episodios por capítulo CIE-10

Cap. CIE-10	Frecuencia	Porcentaje	Nombre clínico del capítulo
O	15.228	16,1769	Embarazo, parto y puerperio.
R	11.661	12,3877	Síntomas, signos y hallazgos anormales.
N	8.876	9,4291	Enfermedades del sistema genitourinario.
C	8.766	9,3123	Tumores.
I	7.015	7,4521	Enfermedades del sistema circulatorio.
K	6.515	6,9210	Enfermedades del sistema digestivo.
J	6.227	6,6150	Enfermedades del sistema respiratorio.
A	4.242	4,5063	Ciertas enfermedades infecciosas y parasitarias.
D	3.723	3,9550	Tumores / sangre y órganos hematopoyéticos.
M	3.262	3,4653	Enfermedades del sistema osteomuscular.
S	2.744	2,9150	Traumatismos y envenenamientos.

Se compararon configuraciones de 5, 10, 20 y 30 componentes SVD en los macrogrupos sometidos a agrupamiento. Esta sensibilidad permitió evaluar simultáneamente varianza acumulada retenida, información no retenida, métricas internas de agrupamiento, concordancia frente a la

configuración final de 10 componentes y tamaño mínimo de grupo.

Tabla 15

Análisis de sensibilidad del número de componentes SVD en el macrogrupo no transmisible

Comp. SVD	Var. acum.	Info. no ret.	Silhouette	Calinski-H.	Davies-B.	Tamaño mín.	Decisión
5	0,2350	0,7650	0,4832	2326,3340	0,8675	1863	Sensibilidad
10	0,3980	0,6020	0,2602	945,5328	1,7162	1864	Configuración final
20	0,5909	0,4091	0,1420	465,4044	1,8862	973	Sensibilidad
30	0,6948	0,3052	0,0726	261,1289	2,6911	755	Sensibilidad

Tabla 16

Análisis de sensibilidad del número de componentes SVD en el macrogrupo infecciosa no COVID

Comp. SVD	Var. acum.	Info. no ret.	Silhouette	Calinski-H.	Davies-B.	Tamaño mín.	Decisión
5	0,5239	0,4761	0,3822	2021,8776	1,2719	1618	Sensibilidad
10	0,6892	0,3108	0,2483	1279,7626	1,5977	1704	Configuración final
20	0,8557	0,1443	0,2503	1114,0008	1,8933	1224	Sensibilidad
30	0,9526	0,0474	0,2222	949,2976	2,0331	1225	Sensibilidad

La elección de 10 componentes SVD no se interpretó como una maximización de la varianza acumulada. Se adoptó como una decisión de compromiso entre reducción de dimensionalidad, conservación parcial de la estructura categórica y posibilidad de obtener grupos interpretables. En NO_TRANSMISIBLE, los 10 componentes conservaron el 39,8 % de la variabilidad de la matriz binaria generada mediante *one-hot encoding*; en INFECCIOSA_NO_COVID, conservaron el 68,9 %. Esta diferencia es coherente con la mayor heterogeneidad diagnóstica y cardinalidad del macrogrupo no transmisible. Las curvas de varianza explicada acumulada se presentan en los Complementos del Capítulo 5, Figuras 16 y 17.

Se evaluaron valores de k entre 2 y 8 en los macrogrupos sometidos a subdivisión algorítmica. En este capítulo, dicha evaluación se presenta como evidencia preliminar para documentar la

preparación del análisis, no como la exposición definitiva de la solución clínica. La selección final de k , la estabilidad de la partición, los nombres interpretativos de los grupos y su caracterización clínica se desarrollan en el Capítulo 7.

Tabla 17

Métricas internas preliminares para evaluación de k en macrogrupos subdivididos

Macrogrupo	k	Silhouette	Calinski-H.	Davies-B.	Tamaño mín.
NO_TRANSMISIBLE	2	0,2017	866,3958	1,3222	1864
NO_TRANSMISIBLE	3	0,2622	961,2903	1,7024	1864
NO_TRANSMISIBLE	4	0,2906	883,3101	1,5360	871
NO_TRANSMISIBLE	5	0,3405	971,6363	1,3809	871
NO_TRANSMISIBLE	6	0,3799	1085,6948	1,2017	871
NO_TRANSMISIBLE	7	0,4272	1291,0122	1,1467	871
NO_TRANSMISIBLE	8	0,4534	1376,2735	1,0855	860
INFECCIOSA_NO_COVID	2	0,1926	1033,0039	1,9458	1704
INFECCIOSA_NO_COVID	3	0,2476	1275,9432	1,5992	1704
INFECCIOSA_NO_COVID	4	0,3135	1354,4285	1,3956	552
INFECCIOSA_NO_COVID	5	0,3417	1488,2610	1,1950	552
INFECCIOSA_NO_COVID	6	0,3690	1542,6023	1,2341	552
INFECCIOSA_NO_COVID	7	0,4059	1627,9181	1,1540	491
INFECCIOSA_NO_COVID	8	0,4200	1566,5846	1,0772	253

La evaluación preliminar mostró que valores mayores de k podían mejorar algunas métricas internas, pero también aumentaban la fragmentación y reducían el tamaño mínimo de los grupos, especialmente en el macrogrupo infeccioso. Esta evidencia funcionó como insumo para la selección definitiva presentada en el Capítulo 7, donde se integran además estabilidad, interpretabilidad clínica, tamaño de grupo y utilidad institucional. Las visualizaciones de las métricas internas según

k se presentan en los Complementos del Capítulo 5, Figuras 18–23.

5.6 Transformación financiera preliminar

La validación final de estabilidad del agrupamiento no se desarrolla en este capítulo, debido a que corresponde a la evaluación definitiva de la solución clínica obtenida. Así, los resultados de estabilidad mediante cambios de semilla, se reportan en el Capítulo 7. En esta sección se conserva únicamente la evidencia preliminar que sustenta la elección de la transformación financiera utilizada posteriormente.

Antes de seleccionar la transformación financiera principal se compararon distintas escalas del ingreso total. La comparación incluyó asimetría, curtosis, comportamiento gráfico, dispersión, diagnóstico Q–Q, heterocedasticidad y desempeño en el modelo ANCOVA. El análisis se realizó sobre ingreso total del episodio; por tanto, los resultados se interpretaron como ingreso facturado o causado, no como costo, utilidad ni rentabilidad.

Tabla 18*Comparación de transformaciones candidatas del ingreso total*

Transformación	Asimetría	Curtosis	Mediana	IQR	R^2_{aj} ANCOVA	Decisión
Ingreso original	9,4582	203,6193	1.734.110	3.789.711	0,3545	Descartada como escala principal por asimetría y valores extremos.
$\log(Y + c)$	0,6921	0,0278	14,3660	1,6616	0,4814	Candidata; mejora asimetría y conserva interpretación proporcional.
$\log(Y + c)$ robusto	0,6921	0,0278	0,0000	1,0000	0,4814	Transformación final; conserva forma logarítmica y estandariza por mediana e IQR.
Yeo–Johnson	0,0629	-0,5668	-0,0581	1,4377	0,4348	Sensibilidad; mejora simetría marginal, pero no supera el desempeño del modelo principal.
Box–Cox	0,0629	-0,5668	-0,0581	1,4377	0,4348	Sensibilidad; comportamiento similar a Yeo–Johnson.
Winsorización + log	0,6595	-0,1667	–	–	0,4780	Sensibilidad; no se adopta como principal para evitar alterar valores extremos observados.

Nota. El R^2_{aj} corresponde al modelo ANCOVA con interacción entre grupo clínico y estancia, ajustado sobre cada transformación candidata del ingreso total.

La escala original presentó asimetría de 9,4582 y curtosis en exceso de 203,6193, evidenciando concentración en valores bajos y presencia de valores extremos. Las transformaciones Yeo–Johnson y Box–Cox corrigieron mejor la asimetría marginal, pero presentaron menor desempeño en el modelo ANCOVA. La transformación logarítmica robusta ofreció el mejor equilibrio entre reducción de valores extremos, interpretabilidad financiera y desempeño modelístico; por ello fue adoptada como transformación final. Las visualizaciones diagnósticas de estas transformaciones se presentan en los Complementos del Capítulo 5,

Figuras 26, 27 y 28. **6. Análisis post hoc de asociaciones**

6.1 Método, criterios y alcance del análisis post hoc

Con el fin de profundizar en las asociaciones bivariadas priorizadas, se realizó una descomposición del estadístico chi-cuadrado para identificar las combinaciones específicas de categorías que aportaban mayor peso al resultado global. Este análisis se aplicó a cuatro cruces: servicio con diagnóstico de egreso Top 50, servicio con fallecimiento, servicio con asignación de enfermería Top 12 y asignación de enfermería Top 12 con EPS Top 12.

Para cada cruce se calcularon frecuencias observadas, frecuencias esperadas, residuales ajustados, contribución absoluta al estadístico chi-cuadrado, contribución porcentual, valor p por celda y valor p ajustado mediante Holm. La frecuencia esperada bajo independencia se definió como:

$$E_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}. \quad (\text{Ecuación 6.1})$$

La contribución de cada celda al estadístico chi-cuadrado fue:

$$C_{ij} = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \quad (\text{Ecuación 6.2})$$

Adicionalmente, se calcularon residuales ajustados:

$$r_{ij}^{aj} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1 - p_{i\cdot})(1 - p_{\cdot j})}}, \quad (\text{Ecuación 6.3})$$

donde $p_{i\cdot}$ corresponde a la proporción marginal de la fila i y $p_{\cdot j}$ a la proporción marginal de la columna j . Un residual ajustado positivo indica sobrerrepresentación respecto a lo esperado bajo independencia; un residual negativo indica subrepresentación. La interpretación se centró en las celdas con mayor contribución porcentual al estadístico global y se mantuvo como descriptiva, no causal (35, 47).

Tabla 19

Cruces priorizados para el análisis post hoc

Variable 1	Variable 2	Criterio de priorización
Servicio	Diagnóstico de egreso Top 50	Asociación de magnitud moderada–alta; permite revisar mezcla diagnóstica por servicio.
Servicio	Fallecimiento SN	Asociación de magnitud moderada–alta; permite revisar el desenlace externo por servicio.
Servicio	Asignación de enfermería Top 12	Asociación fuerte; funciona como control de coherencia entre servicio y ubicación física.
Asignación de enfermería Top 12	EPS Top 12	Asociación administrativa exploratoria; permite identificar patrones descriptivos por pagador.

Tabla 20*Resumen global de los cruces post hoc*

Cruce	<i>n</i>	Dim.	χ^2 global	gl	% esp. <5	Mín. esp.
Servicio vs diagnóstico de egreso Top 50	94.134	5x51	92.088,48	200	19,2157	0,7769
Servicio vs fallecimiento	94.134	5x2	12.059,48	4	0,0000	5,3013
Servicio vs asignación de enfermería Top 12	94.134	5x13	106.072,33	48	0,0000	6,6342
Asignación de enfermería Top 12 vs EPS Top 12	94.134	13x13	50.667,52	144	0,0000	12,9076

El cruce servicio–diagnóstico presentó celdas esperadas menores a 1 y numerosos ceros observados, por lo cual sus valores *p* se interpretaron con cautela. En ese cruce, el análisis post hoc se utilizó principalmente como descomposición descriptiva de patrones clínicos y no como inferencia concluyente por celda.

6.2 Resultados por cruce priorizado

6.2.1 Servicio y diagnóstico de egreso Top 50

El cruce entre servicio y diagnóstico de egreso Top 50 mostró que las mayores contribuciones al chi-cuadrado se concentraron en diagnósticos obstétricos dentro de Ginecobstetricia y en diagnósticos respiratorios o infecciosos dentro de hospitalización pediátrica. También aparecieron celdas con cero observado en hospitalización de adultos para diagnósticos obstétricos. Estas celdas no se interpretaron como errores de calidad del dato, sino como restricciones asistenciales esperables; no obstante, su presencia afecta la lectura inferencial del chi-cuadrado y obliga a interpretar el cruce como descriptivo.

Tabla 21*Celdas con mayor contribución al chi-cuadrado: servicio y diagnóstico de egreso Top 50*

Servicio	Diagnóstico	Obs.	Esp.	Resid. aj.	% contrib.
Ginecobstetricia	Parto por cesárea, sin otra especificación	4.415	738,1	151,90	19,89
Ginecobstetricia	Parto único espontáneo, presentación cefálica	3.355	560,9	131,64	15,11
	...				
Ginecobstetricia	Parto único espontáneo, sin otra especificación	1.765	295,1	94,66	7,95
Ginecobstetricia	Parto por cesárea de emergencia	1.140	190,6	75,82	5,14
Ginecobstetricia	Parto por cesárea electiva	826	138,1	64,43	3,72
Ginecobstetricia	Otros diagnósticos	2.997	8.259,2	-92,04	3,64
Hospitalización pe- diátrica	Estado asmático	600	83,0	61,14	3,49
Ginecobstetricia	Nacido vivo, único	744	124,6	61,06	3,35
Hospitalización adultos	Parto por cesárea, sin otra especificación	0	3.004,6	-99,34	3,26
Hospitalización pe- diátrica	Bronquiolitis aguda, no especificada	433	57,9	53,05	2,64

Nota. Un residual ajustado positivo indica sobrerrepresentación respecto a lo esperado bajo independencia, mientras que un residual ajustado negativo indica subrepresentación. El porcentaje de contribución corresponde al aporte de cada celda al estadístico chi-cuadrado global.

La visualización correspondiente a este cruce se trasladó a los Complementos del Capítulo 6, Figura 29. En conjunto, el cruce se interpretó como control de coherencia clínica entre servicio y mezcla diagnóstica. No debe presentarse como descubrimiento causal, pues parte importante de la asociación obedece a la organización asistencial esperada de la institución.

6.2.2 Servicio y fallecimiento

En el cruce servicio–fallecimiento, las mayores contribuciones al chi-cuadrado se concentraron en las celdas de fallecimiento en UCI adultos y UCI pediátrica. En contraste, Ginecobstetricia y hospitalización pediátrica presentaron menos fallecimientos de los esperados bajo independencia. Esta lectura es descriptiva y no ajusta por edad, diagnóstico, comorbilidad o severidad clínica.

Tabla 22

Celdas con mayor contribución al chi-cuadrado: servicio y fallecimiento

Servicio	Fallecimiento	Obs.	Esp.	Resid. aj.	% contrib.
UCI adultos	Sí	580	33,8	95,71	73,07
UCI pediátrica	Sí	118	5,3	49,56	19,87
Ginecobstetricia	Sí	1	349,1	-20,65	2,88
Hospitalización pediátrica	Sí	23	278,8	-16,64	1,95
UCI adultos	No	946	1.492,2	-95,71	1,66
UCI pediátrica	No	121	233,7	-49,56	0,45
Ginecobstetricia	No	15.737	15.388,9	20,65	0,07
Hospitalización pediátrica	No	12.545	12.289,2	16,64	0,04
Hospitalización adultos	Sí	1.366	1.421,0	-2,61	0,02
Hospitalización adultos	No	62.697	62.642,0	2,61	0,00

Nota. Un residual ajustado positivo indica sobrerrepresentación respecto a lo esperado bajo independencia, mientras que un residual ajustado negativo indica subrepresentación. El porcentaje de contribución corresponde al aporte de cada celda al estadístico chi-cuadrado global.

La visualización correspondiente a este cruce se trasladó a los Complementos del Capítulo 6, Figura 30. El patrón observado es compatible con diferencias de complejidad entre servicios; sin

embargo, el análisis no estima efectos ajustados. Por tanto, el resultado se interpreta únicamente como asociación descriptiva entre servicio y fallecimiento, y no como efecto causal del servicio sobre la mortalidad.

6.2.3 Servicio y asignación de enfermería Top 12

El cruce servicio–asignación de enfermería Top 12 presentó una estructura fuertemente definida. Las mayores contribuciones correspondieron a combinaciones esperables entre servicio y área física. Por ello, este cruce se interpretó como control de coherencia operativa y no como hallazgo novedoso.

Tabla 23

Celdas con mayor contribución al chi-cuadrado: servicio y asignación de enfermería Top 12

Servicio	Asignación de enfermería	Obs.	Esp.	Resid. aj.	% contrib.
Ginecobstetricia	1 UE Hosp. piso 7 - CAL	9.180	1.966,3	190,56	24,95
Hosp. pediátrica	1 UE Hosp. piso 8 - CAL	6.579	1.149,7	180,47	24,17
UCI adultos	Otras asignaciones	1.526	148,9	119,77	12,01
Ginecobstetricia	1 UE Hosp. piso 7 - sala partos	3.541	756,5	113,70	9,66
Hosp. adultos	1 UE Hosp. piso 7 - CAL	1.310	8.004,0	-141,52	5,28
Ginecobstetricia	1 UE Hosp. urgencias CAL P5	2	4.763,2	-90,52	4,49
Hosp. adultos	1 UE Hosp. piso 8 - CAL	1.351	5.860,2	-109,34	3,27
Hosp. adultos	1 UE Hosp. urgencias CAL P5	26.193	19.388,9	103,53	2,25
UCI pediátrica	Otras asignaciones	239	23,3	47,07	1,88
Hosp. adultos	1 UE Hosp. piso 7 - sala partos	953	3.079,5	-69,49	1,38

Nota. Un residual ajustado positivo indica sobrerrepresentación respecto a lo esperado bajo independencia, mientras que un residual ajustado negativo indica subrepresentación. El porcentaje

de contribución corresponde al aporte de cada celda al estadístico chi-cuadrado global.

La visualización correspondiente a este cruce se trasladó a los Complementos del Capítulo 6, Figura 31. La utilidad de este cruce fue verificar la correspondencia descriptiva entre especialidad clínica y ubicación hospitalaria registrada en la base. No se interpretó como validación formal del proceso de ruteo ni como evidencia causal de asignación de camas.

6.2.4 Asignación de enfermería Top 12 y EPS Top 12

El cruce entre asignación de enfermería Top 12 y EPS Top 12 mostró patrones de sobre-representación y subrepresentación en áreas específicas. La mayor contribución al chi-cuadrado correspondió a la asignación 1 UE HOSP PISO 9 - CAL VIP con varias EPS, especialmente Ecopetrol, Colsanitas y otras aseguradoras de baja frecuencia relativa.

Tabla 24

Celdas con mayor contribución al chi-cuadrado: asignación de enfermería Top 12 y EPS Top 12

Asignación de enfermería	EPS Top 12	Obs.	Esp.	Resid. aj.	% contrib.
1 UE Hosp. piso 9 - CAL VIP	Ecopetrol	2.048	265,9	116,10	23,57
1 UE Hosp. piso 9 - CAL VIP	Colsanitas	900	103,6	82,24	12,08
1 UE Hosp. Clinicentro urg. P4	Fiduprevisora	1.441	246,0	80,91	11,46
1 UE Hosp. piso 9 - CAL VIP	Nueva EPS	1.526	5.679,4	-110,67	5,99
1 UE Hosp. piso 9 - CAL VIP	AXA Colpatria	517	67,6	57,34	5,90
1 UE Hosp. piso 9 - CAL VIP	Otras EPS	1.235	303,1	57,01	5,66
1 UE Hosp. piso 9 - CAL VIP	Allianz Seguros	331	38,5	49,38	4,39
1 UE Hosp. Clinicentro urg. P4	Nueva EPS	171	2.109,4	-82,44	3,52
1 UE Hosp. Clinicentro urg. P4	Ecopetrol	453	98,8	36,84	2,51
1 UE Hosp. piso 9 - CAL VIP	Suramericana SA	366	90,9	30,31	1,64

Nota. Un residual ajustado positivo indica sobrerrepresentación respecto a lo esperado bajo

independencia, mientras que un residual ajustado negativo indica subrepresentación. El porcentaje de contribución corresponde al aporte de cada celda al estadístico chi-cuadrado global.

La visualización correspondiente a este cruce se trasladó a los Complementos del Capítulo 6, Figura 32. La asociación observada indica que la distribución de episodios por asignación de enfermería no fue homogénea entre categorías de EPS recodificadas. Sin embargo, este análisis no incorpora información contractual, disponibilidad de camas, reglas de admisión ni criterios administrativos de asignación. Por tanto, el resultado debe entenderse como exploratorio y útil para orientar preguntas institucionales posteriores, no como evidencia de mecanismos contractuales o de ruteo hospitalario.

6.2.4.1 Nota sobre ceros observados y restricciones estructurales.. En los cruces post hoc se identificaron celdas con frecuencia observada igual a cero y frecuencia esperada positiva. Este patrón fue especialmente relevante en el cruce servicio–diagnóstico, donde algunos diagnósticos obstétricos tuvieron frecuencia esperada alta en hospitalización de adultos, pero frecuencia observada igual a cero. Estos casos se interpretaron como restricciones estructurales o asistenciales esperables, no como anomalías automáticas del dato.

La presencia de ceros estructurales puede aumentar la contribución al estadístico chi-cuadrado y elevar el tamaño de efecto global. Por ello, los cruces con ceros estructurales se interpretaron como patrones descriptivos de organización clínica y no como evidencia inferencial suficiente para establecer mecanismos causales.

Tabla 25

Principales celdas con cero observado y frecuencia esperada positiva en servicio y diagnóstico Top 50

Servicio	Diagnóstico	Obs.	Esp.	Resid. aj.	% contrib.
Hospitalización adultos	Parto por cesárea, sin otra especificación	0	3.004,6	-99,34	3,26
Hospitalización adultos	Parto único espontáneo, presentación cefálica	0	2.283,2	-86,09	2,48
Hospitalización adultos	Parto único espontáneo, sin otra especificación	0	1.201,2	-61,90	1,30
Hospitalización adultos	Parto por cesárea de emergencia	0	775,8	-49,58	0,84
Ginecobstetricia	Infección de vías urinarias, no especificado	0	703,4	-29,73	0,76

Nota. Las primeras cuatro filas corresponden a ceros estructurales probables, dado que los diagnósticos obstétricos no se esperarían clínicamente en hospitalización de adultos. La última fila se interpreta como una restricción de mezcla diagnóstica entre servicio y diagnóstico.

6.3 Síntesis interpretativa y alcance del análisis post hoc

Para evitar sobreinterpretaciones, los cruces post hoc se clasificaron según su función dentro del análisis: relaciones estructurales esperables, asociaciones descriptivas clínicas o asociaciones exploratorias administrativas.

Tabla 26*Clasificación interpretativa de los cruces post hoc*

Cruce	Tipo	Advertencia interpretativa
Servicio vs diagnóstico de egreso Top 50	Estructural / clínicamente esperable	No debe presentarse como descubrimiento novedoso; refleja en parte la organización asistencial.
Servicio vs fallecimiento	Descriptivo clínico	No permite inferir causalidad ni efecto del servicio sobre la mortalidad.
Servicio vs asignación de enfermería Top 12	Estructural / operativo esperado	No valida formalmente ruteo interno ni procesos administrativos.
Asignación de enfermería Top 12 vs EPS Top 12	Exploratorio administrativo	No permite concluir rutas contractuales ni decisiones de asignación de camas.

El análisis post hoc permitió identificar las celdas que más contribuyeron a las asociaciones globales priorizadas. En conjunto, los cruces revisados mostraron que una parte importante de los patrones observados corresponde a relaciones estructurales esperables dentro de una institución hospitalaria, especialmente la correspondencia entre servicio, diagnóstico y asignación física. Por esta razón, dichos cruces se interpretaron como controles de coherencia descriptiva y no como hallazgos causales.

El cruce entre servicio y fallecimiento aportó una lectura descriptiva del desenlace externo por tipo de servicio, mientras que el cruce entre asignación de enfermería y EPS Top 12 se consideró exploratorio. En ningún caso los residuales ajustados o las contribuciones al chi-cuadrado permiten inferir por sí solos mecanismos organizacionales, contractuales o clínicos. Los resultados deben entenderse como insumos para priorizar revisiones institucionales posteriores y no como validación

formal de procesos hospitalarios.

7. Agrupamiento clínico no supervisado

7.1 Propósito, unidad de análisis y cohorte

Este capítulo presenta la construcción de grupos clínicos de episodios hospitalarios mediante aprendizaje no supervisado. El propósito no fue predecir mortalidad, estancia ni ingreso, sino obtener una segmentación clínico-diagnóstica interpretable que permitiera organizar la cohorte 2024 y servir como base para los análisis administrativos y financieros posteriores. En consecuencia, la calidad de la solución se evaluó mediante un criterio multicriterio parsimonioso y jerarquizado: (i) coherencia clínica de las variables diagnósticas de entrada, (ii) separación interna suficiente para distinguir perfiles, (iii) tamaños de grupo compatibles con lectura institucional, (iv) estabilidad ante cambios de semilla y remuestreo, y (v) utilidad para caracterización posterior con variables no usadas en el agrupamiento. Estos criterios no se ponderaron como un índice único, porque el objetivo fue construir una segmentación exploratoria defendible y trazable, no optimizar una función supervisada. Por tanto, el agrupamiento se interpreta como herramienta de segmentación clínica institucional y no como clasificación diagnóstica definitiva.

El Capítulo 5 documentó la preparación de la base, las recodificaciones, el análisis bivariado preliminar y las sensibilidades iniciales necesarias para justificar la construcción de grupos clínicos. A partir de ese soporte, este capítulo presenta la solución definitiva de agrupamiento no supervisado, su validación interna, su estabilidad y su caracterización clínica posterior.

La unidad disponible para el análisis fue el **episodio de hospitalización**. Esta definición permitió conservar coherencia entre diagnóstico de egreso, estancia, mortalidad e ingreso del evento analizado. Sin embargo, la base no incluyó un identificador longitudinal de paciente, documento o historia clínica que permitiera evaluar dependencia entre episodios de un mismo individuo, reingresos independientes o trayectorias longitudinales. Como control de trazabilidad, la base final

contenía 26.266 filas y 26.262 identificadores únicos de episodio; cuatro identificadores aparecieron dos veces, para un total de ocho filas vinculadas a identificadores repetidos. Esta diferencia es mínima frente al tamaño total de la cohorte, pero se reporta explícitamente porque esos registros deben auditarse si se requiere unicidad estricta fila–episodio. En consecuencia, las inferencias se restringen al nivel episodio y no se formulan como estimaciones individuales por paciente.

No se utilizaron variables de mortalidad, ingreso, estancia, EPS, servicio hospitalario ni unidad médica para formar los grupos; estas variables se reservaron para análisis *post hoc*. Esta separación evita circularidad metodológica: las variables usadas para interpretar los grupos no fueron usadas por el algoritmo para construirlos.

Tabla 27*Definiciones operativas utilizadas en el capítulo*

Término	Definición operativa
Episodio	Registro hospitalario consolidado asociado a un ingreso y egreso. Es la unidad de análisis.
Macrogrupo clínico	Categoría clínica reconstruida mediante reglas jerárquicas aplicadas al código CIE-10 y al nombre del diagnóstico.
Subcluster algorítmico	Subgrupo obtenido dentro de un macrogrupo mediante la estrategia MiniBatchKMeans–Ward.
Grupo clínico final	Identificador final que combina macrogrupo y subcluster cuando hubo subdivisión.
Mortalidad local	Porcentaje de fallecimientos dentro del grupo; se calcula después de formar los grupos.
Aporte global a muertes	Porcentaje de fallecimientos del grupo sobre el total de fallecimientos de la cohorte.
Validación interna	Evaluación basada en métricas de separación y cohesión del agrupamiento.
Caracterización externa/post hoc	Evaluación con variables no usadas para agrupar, como mortalidad, servicio, unidad médica o ingreso. En este capítulo se usa como caracterización descriptiva, no como prueba predictiva.

Nota. La terminología se fija para evitar el uso indistinto de *cluster*, grupo, subgrupo, segmento y macrogrupo.

Tabla 28
Verificación de unicidad del identificador de episodio en la base final

Indicador	Valor
Filas de la base final	26.266
Identificadores únicos de episodio	26.262
Filas pertenecientes a episodios con identificador repetido	8
Identificadores de episodio con más de una fila	4

Nota. El control evidencia una discrepancia mínima entre filas y episodios únicos. La base disponible no permite evaluar dependencia longitudinal entre episodios de un mismo paciente. Los cuatro identificadores repetidos se reportan como registros a auditar si se requiere unicidad estricta fila–episodio.

7.1.1 Selección de la cohorte de análisis

La base contenía episodios hospitalarios del periodo 2021–2024. La distribución anual se presenta en la Tabla 29. Esta tabla describe disponibilidad de datos, pero no demuestra representatividad estadística del año 2024 frente a los demás años.

Tabla 29
Distribución de episodios por año de análisis

Año de análisis	Número de episodios
2021	19.885
2022	23.582
2023	24.387
2024	26.280

Nota. El volumen anual no se usa como prueba de representatividad. La decisión de usar 2024 obedece a actualidad institucional y pertinencia operativa.

La cohorte principal fue 2024. Esta elección se sustentó en la actualidad institucional de la información y en la recomendación clínico-operativa del asesor médico del proyecto. Esta recomendación se interpreta como criterio para focalizar el análisis, no como validación clínica externa formal ni como demostración de representatividad estadística del comportamiento institucional de 2021–2024. En consecuencia, el capítulo no sostiene que 2024 sea “mejor”, “más limpio” o suficientemente libre de sesgos pospandemia; lo usa como cohorte institucional vigente para segmentación y análisis posterior. No se contó con información documental suficiente para evaluar cambios interanuales en políticas de codificación, sistemas de información hospitalaria, prácticas de referencia, cobertura de aseguramiento o perfil de usuarios atendidos. Por tanto, la generalización a otros años debe hacerse con cautela y los hallazgos deben leerse como propios de la cohorte 2024. La trazabilidad administrativa de la recomendación se conserva en los apéndices para no sobrecargar el cuerpo principal.

7.2 Reclasificación clínica, discrepancia y sensibilidad COVID-19

Antes del agrupamiento se reconstruyó la clasificación clínica de los diagnósticos para reducir inconsistencias entre la gran causa original y el diagnóstico de egreso. La reclasificación usó tres fuentes: código CIE-10, nombre del diagnóstico y reglas clínicas jerárquicas. El procedimiento asignó cada episodio a una única categoría: COVID-19, materna-obstétrica, perinatal-neonatal, lesiones, infecciosa no COVID, mal definidas o inespecíficas, o no transmisible.

Las reglas fueron implementadas computacionalmente y revisadas en el marco de la asesoría médica del proyecto, con participación del médico Julián Navas como asesor médico. Esta revisión permitió depurar criterios operativos, pero no constituye validación clínica externa formal, porque no se dispone de una muestra independiente clasificada manualmente por evaluadores clínicos. Por ello, esta etapa no reporta kappa de Cohen, precisión, sensibilidad, *recall* ni F1. La reclasificación debe entenderse como control reproducible de coherencia diagnóstica basado en CIE-10 y texto diagnóstico, no como auditoría clínica individual ni como sustituto de la historia clínica.

Tabla 30*Orden jerárquico de reclasificación clínica*

Orden	Categoría asignada	Criterio operativo general
1	COVID-19	Códigos U07.1, U07.2 o expresiones asociadas a COVID-19, SARS-CoV-2 o coronavirus 2019.
2	Materna-obstétrica	Capítulo O, códigos Z34–Z39, Z37 o expresiones asociadas a embarazo, parto, cesárea, aborto o preeclampsia.
3	Perinatal-neonatal	Capítulo P o expresiones asociadas a recién nacido, neonatalidad, prematuridad, bajo peso al nacer o inmadurez extrema.
4	Lesiones	Capítulos S, T, V, W, X, Y o expresiones asociadas a trauma, fractura, accidente, quemadura, intoxicación o envenenamiento.
5	Infecciosa no COVID	Capítulos A y B, códigos y patrones de sepsis, infecciones respiratorias, urinarias y sistema nervioso central, excluyendo COVID-19.
6	Mal definidas o inespecíficas	Capítulo R, diagnósticos vacíos o expresiones asociadas a signos, síntomas, observación por sospecha o diagnóstico pendiente.
7	No transmisible	Categoría final para episodios que no cumplieron las reglas anteriores.

Tabla 31*Códigos y capítulos CIE-10 usados en la reclasificación clínica*

Categoría	Códigos o capítulos CIE-10
COVID-19	U07.1, U07.2.
Materna-obstétrica	Capítulo O; códigos Z34–Z39; Z37.
Perinatal-neonatal	Capítulo P.
Lesiones	Capítulos S, T, V, W, X, Y.
Infecciosa no COVID	Capítulos A y B; A40–A41; R57.2; G00–G09; J09–J22; N10–N16; L00–L08.
Mal definidas	Capítulo R; código diagnóstico vacío o nombre diagnóstico vacío.

Nota. La tabla separa los criterios codificados para evitar una tabla horizontal excesiva. Los patrones textuales complementarios se presentan en la Tabla 32.

Tabla 32*Patrones textuales principales usados en la reclasificación clínica*

Categoría	Patrones textuales principales
COVID-19	COVID, SARS-CoV-2, coronavirus 2019, infección por coronavirus.
Materna-obstétrica	Cesárea, parto, trabajo de parto, puerperio, embarazo, gestación, aborto, obstétrico, preeclampsia, eclampsia.
Perinatal-neonatal	Recién nacido, neonatal, prematuro, prematuridad, bajo peso al nacer, distrés respiratorio del recién nacido, inmadurez extrema, retinopatía de la prematuridad.
Lesiones	Trauma, traumatismo, fractura, herida, accidente, quemadura, intoxicación, envenenamiento.
Infecciosa no COVID	Sepsis, septicemia, neumonía, infección urinaria, dengue, tuberculosis, VIH, SIDA, meningitis, celulitis, bronquiolitis, pielonefritis, cistitis, erisipela, fascitis, malaria, leptospirosis.
Exclusiones infecciosas explícitas	Apendicitis, cálculo vesicular o biliar, colecistitis, colangitis, obesidad, contacto o pesquisa para tuberculosis.
Mal definidas o inespecíficas	Signos y síntomas, hallazgos anormales, observación por sospecha, dolor abdominal, fiebre no especificada, síndrome febril, malestar general, diagnóstico pendiente o por confirmar.

Nota. La categoría mal definida no implica ausencia de relevancia clínica; identifica episodios cuyo diagnóstico de egreso funciona como signo, síntoma o condición inespecífica según la información codificada disponible. No se reclasificaron estos diagnósticos hacia etiologías específicas porque ello requeriría historia clínica, evolución médica o resultados paraclínicos no disponibles en la base.

La categoría mal definidas o inespecíficas se conserva como categoría residual controlada, no como juicio de baja importancia clínica. Diagnósticos como fiebre, disnea, síncope, dolor abdominal o dolor torácico pueden representar cuadros clínicos relevantes; sin embargo, con la información disponible no es metodológicamente válido reasignarlos a una causa específica sin evidencia clínica adicional. Esta decisión reduce el riesgo de reclasificación especulativa y hace explícito el posible sesgo de información asociado a diagnósticos codificados como signos, síntomas o condiciones inespecíficas.

Tabla 33*Verificación de asignación completa posterior a la reclasificación*

Indicador	Valor
Episodios analizados sin COVID-19	26.266
Episodios con categoría clínica reconstruida	26.266
Episodios sin categoría asignada	0

Nota. La asignación completa se obtiene por la estructura jerárquica de reglas y por la categoría residual no transmisible. La tabla no valida clínicamente cada reclasificación individual; solo verifica que no quedaron episodios sin categoría final.

7.2.1 Discrepancia entre clasificación original y reconstruida

Se construyó la variable `flag_discrepancia_clinica` para identificar episodios en los que la clasificación original no coincidía con la clasificación reconstruida. Esta variable no se usó para agrupar, porque representa un indicador de consistencia y trazabilidad de la codificación, no una característica clínica primaria del episodio. Su uso se reservó para auditoría *post hoc*: distribución por macrogrupo, distribución por grupo clínico final y mortalidad estratificada según presencia de discrepancia.

Tabla 34*Discrepancia clínica por macrogrupo en la base final sin COVID-19*

Macrogrupo	n	Discrepancias	Discrepancia %
LESIONES	1.133	1.035	91,35
infecciosa no COVID	5.330	1.977	37,09
materna-obstétrica	4.229	272	6,43
no transmisible	11.638	544	4,67
mal definidas o inespecíficas	3.328	54	1,62
perinatal-neonatal	608	3	0,49

Nota. La discrepancia identifica diferencias entre clasificación original y reconstruida. No equivale automáticamente a error clínico individual.

Las discrepancias se concentraron especialmente en LESIONES y infecciosa no COVID. Este hallazgo respalda conservar la variable como control de codificación y como señal para revisar patrones de reclasificación, pero no implica por sí mismo error clínico individual. Al comparar mortalidad entre episodios con y sin discrepancia, no se observó una diferencia estadísticamente significativa: 1,60 % en episodios con discrepancia frente a 1,82 % en episodios sin discrepancia, prueba chi-cuadrado $p = 0,3334$. Por tanto, con la información disponible no hay evidencia de que la discrepancia clínica se asocie a mayor mortalidad cruda. El detalle por grupo clínico final se mantiene en el apéndice para documentar concentración de discrepancias sin sobrecargar el cuerpo principal.

7.2.2 Sensibilidad por exclusión de COVID-19

La cohorte 2024 tuvo 26.280 episodios. La categoría COVID-19 concentró 14 episodios, equivalentes al 0,05 % de la cohorte, y no registró fallecimientos. Después de excluir COVID-19, la base analítica principal quedó en 26.266 episodios y 469 fallecimientos.

Tabla 35*Sensibilidad descriptiva por exclusión de COVID-19*

Cohorte	Episodios	Fallecimientos	Mortalidad cruda %
2024 con COVID-19	26.280	469	1,7846
2024 sin COVID-19	26.266	469	1,7856
Solo COVID-19	14	0	0,0000

Nota. La exclusión no cambia el numerador de fallecimientos y modifica marginalmente el denominador. No reemplaza una evaluación algorítmica completa de inclusión/exclusión de esos 14 casos.

La exclusión se considera de bajo impacto descriptivo sobre la mortalidad global, porque no modifica el numerador de fallecimientos y cambia de manera marginal el denominador. Sin embargo, esta tabla no prueba estabilidad estructural de los grupos vecinos. Una evaluación completa requeriría recalcular la solución incluyendo COVID-19 y comparar la partición resultante. Por tanto, el efecto exacto de esos 14 casos sobre la estructura algorítmica de grupos infecciosos o mal definidos se mantiene como limitación explícita, no como supuesto resuelto.

7.3 Preparación de variables y estrategia de agrupamiento

La selección de variables se hizo por macrogrupo. En no transmisible la versión final no incluyó sexo, con el fin de priorizar estructura diagnóstica y edad sobre una separación demográfica en el macrogrupo de mayor heterogeneidad clínica. Esta decisión no implica irrelevancia clínica del sexo; por el contrario, el sexo se conserva para análisis *post hoc* y sí se incluyó en otros macrogrupos donde podía capturar perfiles clínico-demográficos. La consecuencia metodológica es que las diferencias por sexo deben interpretarse como parte de la caracterización posterior y no como desenlace causal.

Tabla 36*Variables disponibles/preparadas por macrogrupo para el agrupamiento y la caracterización*

Macrogrupo	Variables utilizadas para el agrupamiento
No transmisible	Grupo de edad, capítulo CIE-10, prefijo CIE-3.
Infecciosa no COVID	Grupo de edad, sexo, capítulo CIE-10, prefijo CIE-3.
materna-obstétrica	Grupo de edad, subtipo obstétrico, prefijo CIE-3.
perinatal-neonatal	Grupo de edad, capítulo CIE-10, prefijo CIE-3.
LESIONES	Grupo de edad, sexo, capítulo CIE-10, prefijo CIE-3.
mal definidas o inespecíficas	Grupo de edad, sexo, capítulo CIE-10, prefijo CIE-3.

Nota. La mortalidad, estancia, ingreso, EPS, servicio y unidad médica no se usaron para formar los grupos.

Las variables categóricas se transformaron con codificación *one-hot*. La edad se usó por grupos etarios como variable categórica; aunque estos grupos tienen un orden natural, no se impuso una distancia ordinal para evitar asumir linealidad entre intervalos clínicamente heterogéneos. Luego se aplicó *TruncatedSVD* a la matriz dispersa para obtener una representación continua de menor dimensión (7). Esta decisión fue apropiada para una matriz categórica dispersa porque reduce dimensionalidad y costo computacional sin requerir densificar la matriz completa. No obstante, las dimensiones SVD no tienen interpretación clínica directa; por ello, la interpretación de los grupos se hizo posteriormente con diagnósticos, edad, sexo, servicio, unidad médica y mortalidad, no con los componentes latentes de manera aislada.

Tabla 37*Resumen de codificación y reducción SVD en macrogrupos subdivididos*

Macrogrupo	<i>n</i>	OHE	SVD	Var. acum.
No transmisible	11.638	672	10	0,3980
Infecciosa no COVID	5.330	136	10	0,6892

Nota. OHE: columnas generadas por codificación *one-hot*; SVD: componentes retenidos; Var. acum.: varianza acumulada. La menor varianza acumulada de no transmisible refleja mayor heterogeneidad diagnóstica. La reducción SVD mejora factibilidad computacional, pero reduce interpretabilidad directa de cada dimensión.

La sensibilidad al número de componentes mostró que la solución no es completamente invariante a esta decisión. Con 5 componentes se obtuvieron siluetas mayores, pero menor varianza acumulada; con 20 y 30 componentes aumentó la varianza retenida pero disminuyó la similitud frente a la configuración final de 10 componentes, especialmente en no transmisible. Por ello, la elección de 10 componentes se documenta como configuración principal parsimoniosa, no como número óptimo universal. El detalle se presenta en la Tabla 64.

7.3.1 Estrategia de agrupamiento

La estrategia combinó microagrupamiento con *MiniBatchKMeans* y agrupamiento jerárquico de Ward sobre centroides. *MiniBatchKMeans* permite resumir grandes conjuntos de datos mediante prototipos con costo computacional menor que K-means completo (9). Ward se aplicó sobre centroides en el espacio SVD usando distancia euclidiana, coherente con el criterio de mínima varianza intragrupo del método (10). La elección de distancia euclidiana se justifica por operar sobre una representación continua reducida, no sobre las variables categóricas originales. Aun así, se reconoce que esta distancia no equivale directamente a similitudes categóricas como Jaccard o coseno; por esa razón, la solución se validó mediante estabilidad y caracterización clínica posterior. El procedimiento fue:

$$\text{episodios} \longrightarrow \text{microgrupos} \longrightarrow \text{centroides} \longrightarrow \text{grupos finales.} \quad (7.1)$$

Tabla 38*Parámetros computacionales de la estrategia de agrupamiento*

Elemento	Especificación
Representación de entrada	Matriz <i>one-hot</i> reducida mediante <i>TruncatedSVD</i> .
Componentes SVD	10 en la solución principal.
Microagrupamiento no supervisado	<i>MiniBatchKMeans</i> .
Regla de microgrupos	$n_{micro} = \min(120, \max(20, \lfloor n/30 \rfloor))$. En los macrogrupos grandes el máximo fue 120.
batch_size	2.048.
n_init	10.
Semilla base	42.
Agrupamiento final	Jerárquico aglomerativo con método de Ward.
Métrica	Distancia euclidiana sobre componentes SVD.
Asignación final	Cada episodio heredó el grupo del centroide de su microcluster.

Nota. La agregación por microgrupos introduce aproximación; por ello se evaluó estabilidad por semillas, remuestreo y sensibilidad SVD.

El paso de episodios a microgrupos introduce una aproximación porque cada centroide resume varios episodios en el espacio SVD. Para cuantificar esta pérdida de forma descriptiva, se calculó la distancia de cada episodio a su centroide de microcluster. La distancia media global fue 0,030 en no transmisible y 0,056 en infecciosa no COVID; la inercia media por episodio fue 0,0013 y 0,0098, respectivamente. Estos valores sugieren una agregación más compacta en no transmisible y mayor heterogeneidad local en el macrogrupo infeccioso. El detalle se reporta en la tabla 88.

La solución sigue siendo una segmentación descriptiva escalable, no la recuperación exacta de una estructura latente única.

Métodos alternativos como DBSCAN o modelos de mezclas gaussianas no se adoptaron como estrategia principal porque el objetivo era obtener grupos institucionalmente interpretables, de tamaño operativo y trazables sobre una matriz categórica dispersa; además, los métodos basados en densidad pueden ser inestables en espacios de alta dimensionalidad y las mezclas gaussianas imponen supuestos de distribución continua poco alineados con variables *one-hot*. Como control exploratorio adicional, se comparó la solución final con K-means directo, Ward directo o muestral y DBSCAN exploratorio en el espacio SVD. Estos resultados se ubican en la tabla 91; su función es evaluar consistencia metodológica, no reemplazar la solución principal.

7.4 Selección del número de grupos y solución final

Esta sección integra la evaluación del número de grupos y la presentación de la solución final. La finalidad es mantener en un mismo bloque la decisión metodológica sobre k y sus consecuencias directas sobre la partición clínica obtenida.

La selección de k se evaluó mediante coeficiente de silueta, Calinski–Harabasz, Davies–Bouldin y tamaño mínimo observado de grupo (4, 13, 14). Ninguna métrica se usó de forma automática ni aislada. La decisión final siguió una regla multicriterio: primero se descartaron particiones con fragmentación clínica poco útil; después se verificó separación interna; luego se revisó estabilidad y coherencia diagnóstica; finalmente se eligió la solución más parsimoniosa que conservara perfiles interpretables para análisis posterior.

Tabla 39*Evaluación de k en no transmisible*

k	Silueta	Calinski–Harabasz	Davies–Bouldin	Tamaño mínimo
2	0,202	866,40	1,322	1.864
3	0,262	961,29	1,702	1.864
4	0,291	883,31	1,536	871
5	0,340	971,64	1,381	871

Nota. $k = 3$ no maximiza todas las métricas internas; se seleccionó como solución multicriterio parsimoniosa porque conserva tamaños robustos, lectura clínica y trazabilidad operativa.

Tabla 40*Evaluación de k en infecciosa no COVID*

k	Silueta	Calinski–Harabasz	Davies–Bouldin	Tamaño mínimo
2	0,193	1.033,00	1,946	1.704
3	0,248	1.275,94	1,599	1.704
4	0,314	1.354,43	1,396	552
5	0,342	1.488,26	1,195	552
6	0,369	1.542,60	1,234	552
7	0,406	1.627,92	1,154	491
8	0,420	1.566,58	1,077	253

Nota. Las métricas internas mejoran con soluciones más fragmentadas; $k = 8$ produce un grupo mínimo de 253 episodios. Se seleccionó $k = 3$ porque preserva la interpretación clínica de perfiles amplios y reduce fragmentación para uso institucional.

Los valores de silueta entre 0,20 y 0,42 indican solapamiento clínico apreciable. Esto

no invalida la segmentación, pero delimita su alcance: los grupos no deben interpretarse como clases clínicas mutuamente excluyentes con fronteras tajantes, sino como perfiles dominantes dentro de una cohorte hospitalaria heterogénea. La solución $k = 3$ se seleccionó como partición parsimoniosa para los dos macrogrupos de mayor tamaño y relevancia analítica. Los macrogrupos LESIONES, mal definidas o inespecíficas, materna-obstétrica y perinatal-neonatal fueron evaluados exploratoriamente; la decisión de mantenerlos como grupos únicos se sustentó en la combinación de tamaño, finalidad descriptiva, estabilidad operativa y bajo beneficio interpretativo de introducir subdivisiones adicionales. La evaluación resumida se reporta en la Tabla 66. Para estos macrogrupos no se reportan ARI ni NMI de una partición interna final, porque la solución principal los conserva como grupos únicos; por tanto, no existe un subagrupamiento final de etiquetas contra la cual evaluar estabilidad.

7.5 Caracterización de grupos y análisis de mortalidad

Esta sección resume la composición de los grupos clínicos finales y sus desenlaces externos. La mortalidad se presenta como análisis *post hoc*; no participó en el algoritmo de agrupamiento.

La solución final contiene 10 grupos clínicos: tres subgrupos de no transmisible, tres de infecciosa no COVID y un grupo único para cada macrogrupo restante. La nomenclatura computacional C01, C02 y C03 se conserva en la base para garantizar trazabilidad y reproducibilidad; sin embargo, el capítulo no interpreta esos códigos como nombres clínicos. Para mejorar la lectura académica y responder a la observación sobre nomenclatura poco informativa, cada grupo recibe una **etiqueta clínica interpretativa** construida a partir de criterios observados después del agrupamiento: macrogrupo, edad, diagnóstico dominante, sexo cuando participa en el algoritmo, mortalidad cruda o ajustada, aporte absoluto a muertes y concentración de discrepancia clínica. Estas etiquetas son descriptivas; no convierten los grupos en diagnósticos definitivos ni reemplazan los códigos computacionales originales.

Tabla 41*Nombres interpretativos de los grupos clínicos finales*

Grupo clínico final	Nombre interpretativo utilizado en el capítulo
No transmisible C01	No transmisible digestivo-hemorrágico de baja mortalidad.
No transmisible C02	No transmisible de adulto mayor cardiovascular-neurológico de alta mortalidad.
No transmisible C03	No transmisible heterogéneo de alto volumen y aporte al total de fallecimientos.
Infecciosa no COVID C01	Infecciosa no COVID masculina, predominantemente dengue/gastroenteritis/neumonía, con mayor mortalidad ajustada por edad y sexo.
Infecciosa no COVID C02	Infecciosa no COVID femenina, predominantemente dengue/gastroenteritis/neumonía, con mortalidad intermedia ajustada por edad y sexo.
Infecciosa no COVID C03	Infecciosa no COVID urinaria y de tejidos blandos, con menor mortalidad cruda y alta reclasificación.
Mal definidas C01	Sintomático inespecífico o mal definido, prioritario para auditoría de codificación.
Lesiones C01	Lesiones y trauma hospitalario.
Materna-obstétrica C01	Materna-obstétrica sin muertes observadas en la cohorte.
Perinatal-neonatal C01	Perinatal-neonatal de bajo volumen y estimación de mortalidad imprecisa.

Nota. Los nombres son etiquetas interpretativas post hoc. La base conserva los identificadores computacionales originales para trazabilidad, pero en el cuerpo del capítulo se utilizan nombres extendidos para facilitar la lectura clínica y evitar abreviaturas poco informativas.

Tabla 42*Mortalidad cruda por grupo clínico final con intervalos de confianza del 95 %*

Grupo clínico final	<i>n</i>	Fallec.	Mort. %	IC95 % inf.	IC95 % sup.	Aporte muertes %
No transmisible C02	2.660	107	4,02	3,34	4,84	22,81
Infecciosa no COVID C01	1.704	56	3,29	2,54	4,24	11,94
Infecciosa no COVID C02	1.803	43	2,38	1,78	3,20	9,17
No transmisible C03	7.114	139	1,95	1,66	2,30	29,64
Mal definidas C01	3.328	59	1,77	1,38	2,28	12,58
Lesiones C01	1.133	19	1,68	1,08	2,60	4,05
Infecciosa no COVID C03	1.823	22	1,21	0,80	1,82	4,69
Perinatal-neonatal C01	608	6	0,99	0,45	2,14	1,28
No transmisible C01	1.864	18	0,97	0,61	1,52	3,84
Materna-obstétrica C01	4.229	0	0,00	0,00	0,09	0,00

Nota. IC95 % binomial. La mortalidad no se usó para formar grupos. El 0 % observado en materna-obstétrica no debe interpretarse como ausencia de riesgo clínico.

La tabla distingue riesgo local y carga absoluta. No transmisible C02, etiquetado como no transmisible de adulto mayor cardiovascular-neurológico, tuvo la mayor mortalidad cruda local, 4,02 % (IC95 %: 3,34–4,84). No transmisible C03, etiquetado como no transmisible heterogéneo de alto volumen, aportó la mayor proporción de fallecimientos totales, 29,64 %, por su tamaño poblacional. En Perinatal-neonatal C01, etiquetado como perinatal-neonatal de bajo volumen, la mortalidad observada fue 0,99 %, pero el intervalo fue amplio (0,45–2,14), lo que confirma la necesidad de reportar incertidumbre en grupos pequeños.

Tabla 43*Caracterización de grupos priorizados: mortalidad y volumen*

Grupo	<i>n</i>	Fallec.	Mort. %	Mediana edad
No transmisible C02	2.660	107	4,02	79
Infecciosa no COVID C01	1.704	56	3,29	21
Infecciosa no COVID C02	1.803	43	2,38	27
No transmisible C03	7.114	139	1,95	53
Mal definidas C01	3.328	59	1,77	62

Nota. Se muestran grupos priorizados con criterios cuantitativos: mayor mortalidad cruda local, mayor aporte absoluto a muertes o concentración relevante de diagnósticos inespecíficos para auditoría.

Tabla 44*Caracterización de grupos priorizados: distribución de edad*

Grupo	Edad media	DE	P25	P75
No transmisible C02	73,30	17,36	65	85
Infecciosa no COVID C01	34,53	31,21	7	65
Infecciosa no COVID C02	37,65	31,68	10	69
No transmisible C03	47,25	24,16	29	67
Mal definidas C01	54,74	26,53	35	76

Nota. La división en dos tablas mantiene lectura vertical y evita comprimir excesivamente la información de edad.

No transmisible C02 recibe la etiqueta *no transmisible de adulto mayor cardiovascular-neurológico de alta mortalidad*. Esta denominación se sustenta en su edad avanzada, con mediana de 79 años y rango intercuartílico 65–85, y en sus diagnósticos dominantes: enfermedad cerebrovascular no especificada (14,02 %), insuficiencia cardiaca congestiva (5,86 %), hipertensión esencial

(5,45 %) y EPOC con exacerbación aguda (3,99 %). La concentración en hospitalización adultos fue 95,08 %. La etiqueta sintetiza edad, diagnóstico dominante y mortalidad cruda; no pretende establecer una entidad clínica única.

No transmisible C03 recibe la etiqueta *no transmisible heterogéneo de alto volumen y aporte al total de fallecimientos*. Su mortalidad local fue menor que la de No transmisible C02, pero su tamaño produjo el mayor aporte absoluto a muertes. Los diagnósticos principales fueron cólico renal no especificado (2,24 %), estado asmático (2,16 %), epilepsia no especificada (1,98 %), gonartrosis primaria bilateral (1,56 %) y lumbago no especificado (1,50 %). La etiqueta prioriza volumen, heterogeneidad diagnóstica y utilidad para el análisis financiero posterior.

En los grupos infecciosos, Infecciosa no COVID C01 e Infecciosa no COVID C02 reciben etiquetas clínico-demográficas: *infeccioso no COVID masculino de mayor mortalidad ajustada por edad y sexo* e *infeccioso no COVID femenino de mortalidad intermedia ajustada por edad y sexo*. Ambos estuvieron dominados por dengue clásico, gastroenteritis infecciosa y neumonías. La variable sexo participó en este macrogrupo y la solución separó C01 como 100 % hombres y C02 como 100 % mujeres. Por tanto, estos grupos no deben interpretarse como entidades etiológicas distintas únicamente por diagnóstico, sino como perfiles clínico-demográficos dentro del macrogrupo infeccioso. Infecciosa no COVID C03 se interpreta como un perfil infeccioso urinario y de tejidos blandos de menor mortalidad cruda y alta reclasificación, de acuerdo con la concentración de discrepancia observada en el apéndice.

7.5.1 Mortalidad: análisis crudo, pruebas de comparación y ajuste por edad/sexo

La mortalidad global de la cohorte final sin COVID-19 fue:

$$\text{Mortalidad global} = \left(\frac{469}{26,266} \right) \times 100 = 1,79 \%. \quad (7.2)$$

Esta cifra se usó como referencia descriptiva, pero las comparaciones formales se hicieron mediante tablas 2x2 grupo versus resto de cohorte, con ajuste de Holm para múltiples compara-

ciones. A nivel de grupo final, No transmisible C02 tuvo odds crudo de mortalidad 2,69 veces el resto de la cohorte ($p_{Holm} < 0,001$) e Infecciosa no COVID C01 tuvo odds crudo 1,99 veces el resto ($p_{Holm} < 0,001$). No transmisible C01 presentó odds crudo menor que el resto ($OR=0,52$; $p_{Holm} = 0,0388$). Las demás diferencias no fueron significativas después del ajuste de Holm.

Tabla 45

Comparación de mortalidad cruda de cada grupo frente al resto de la cohorte

Grupo clínico final	OR crudo	p	p_{Holm}	Lectura
No transmisible C02	2,69	< 0,001	< 0,001	Mayor mortalidad cruda que el resto.
Infecciosa no COVID C01	1,99	< 0,001	< 0,001	Mayor mortalidad cruda que el resto.
No transmisible C01	0,52	0,0055	0,0388	Menor mortalidad cruda que el resto.
Infecciosa no COVID C02	1,38	0,0464	0,2787	No significativo tras Holm.
Infecciosa no COVID C03	0,66	0,0531	0,2787	No significativo tras Holm.
Perinatal-neonatal C01	0,54	0,1324	0,5296	No significativo tras Holm.
No transmisible C03	1,14	0,2093	0,6280	No significativo tras Holm.
Mal definidas C01	0,99	0,9526	1,0000	No significativo.
Lesiones C01	0,94	0,7778	1,0000	No significativo.
Materna-obstétrica C01	0,00	< 0,001	< 0,001	Sin muertes observadas; evento raro.

Nota. OR: odds ratio crudo grupo versus resto. La fila materna-obstétrica se debe interpretar con cautela por ausencia de eventos observados.

Para atender la posible confusión por edad y sexo, se estimó una regresión logística penalizada L2 y se aplicó estandarización marginal. La estandarización marginal estima la mortalidad

esperada de cada grupo bajo una estructura común de edad y sexo, lo que permite comparar perfiles reduciendo diferencias demográficas básicas. Este ajuste no incluye comorbilidades, porque la base disponible para este capítulo y el ejemplo de base hospitalaria no incluyen diagnósticos secundarios ni índices tipo Charlson o Elixhauser. Por tanto, el ajuste reduce parcialmente la confusión demográfica, pero no controla severidad clínica, multimorbilidad ni complejidad asistencial.

Tabla 46

Mortalidad ajustada por edad y sexo mediante estandarización marginal

Grupo clínico final	Mort. ajustada %	IC95 % inf.	IC95 % sup.	Mort. cruda %	Dif. pp
Perinatal-neonatal C01	4,35	1,57	8,90	0,99	3,36
Infecciosa no COVID C01	4,07	3,07	5,13	3,29	0,78
Infecciosa no COVID C02	2,67	1,96	3,35	2,38	0,28
No transmisible C02	2,14	1,74	2,59	4,02	-1,88
No transmisible C03	2,08	1,75	2,43	1,95	0,12
Mal definidas C01	1,42	1,08	1,80	1,77	-0,35
Lesiones C01	1,40	0,84	2,04	1,68	-0,27
Infecciosa no COVID C03	1,08	0,67	1,45	1,21	-0,12
No transmisible C01	0,87	0,58	1,28	0,97	-0,09
Materna-obstétrica C01	0,11	0,10	0,17	0,00	0,11

Nota. Estimación mediante regresión logística L2 y estandarización marginal con 200 remuestreos válidos.

La estimación para grupos con estructura etaria extrema, como perinatal-neonatal, es un contrafactual estadístico y debe interpretarse con cautela.

El ajuste cambia sustancialmente la lectura de No transmisible C02: su mortalidad cruda de 4,02 % disminuye a 2,14 % al estandarizar por edad y sexo, lo que indica que parte importante

de su mortalidad cruda se explica por su estructura etaria avanzada. En contraste, Infecciosa no COVID C01 mantiene una mortalidad ajustada solo por edad y sexo elevada (4,07 %; IC95 %: 3,07–5,13), por lo que conserva relevancia como grupo prioritario según el criterio de riesgo ajustado. La estimación ajustada de Perinatal-neonatal C01 debe interpretarse con especial prudencia por su estructura etaria extrema y por la amplitud del intervalo de confianza. La edad tuvo asociación positiva con mortalidad en el modelo logístico: OR=1,034 por año, IC95 % 1,029–1,039, $p < 0,001$.

7.6 Estabilidad, reproducibilidad, calidad del dato y controles técnicos

Esta sección reúne los controles de robustez y calidad del dato que respaldan el alcance de la segmentación. Primero se reporta la reproducibilidad de la solución y luego los controles técnicos de datos faltantes, diagnósticos raros, plausibilidad clínica y valores atípicos.

La estabilidad se evaluó con 30 corridas por cambio de semilla y 30 remuestreos aleatorios del 80 % para los dos macrogrupos subdivididos. Cada partición se comparó contra la solución base usando ARI y NMI (15, 16). Este procedimiento evalúa sensibilidad frente a inicialización y reducción de muestra, pero no corresponde a una validación cruzada *k-fold*; por tanto, sus resultados se interpretan como evidencia de reproducibilidad interna y no como validación predictiva formal.

Tabla 47*Estabilidad de la solución mediante ARI en 30 corridas*

Macrogrupo	Validación	Corridas	<i>n</i> prom.	Media	DE	Mín.
infecciosa no COVID	Semilla	30	5.330	0,6438	0,1944	0,2180
infecciosa no COVID	Remuestreo 80 %	30	4.264	0,5615	0,1826	0,2662
no transmisible	Semilla	30	11.638	0,7048	0,1446	0,3520
no transmisible	Remuestreo 80 %	30	9.310	0,6749	0,2314	0,2646

Nota. ARI cercano a 1 indica mayor semejanza con la solución base. Los mínimos muestran escenarios de sensibilidad que deben considerarse en la interpretación.

Tabla 48*Estabilidad de la solución mediante NMI en 30 corridas*

Macrogrupo	Validación	Corridas	Media	Mín.	Interpretación
infecciosa no COVID	Semilla	30	0,6690	0,2985	Moderada y variable.
infecciosa no COVID	Remuestreo 80 %	30	0,6053	0,3399	Moderada; sensible al remuestreo.
no transmisible	Semilla	30	0,7439	0,5690	Moderada-alta en promedio.
no transmisible	Remuestreo 80 %	30	0,7414	0,3963	Moderada-alta en promedio, con mínimos sensibles.

Nota. NMI cercano a 1 indica mayor semejanza con la solución base. La tabla separa NMI de ARI para conservar formato vertical y evitar tablas excesivamente comprimidas.

no transmisible presentó estabilidad promedio mayor que infecciosa no COVID. Sin em-

bargo, la estabilidad debe describirse como promedio moderada, con sensibilidad en escenarios mínimos, porque los valores mínimos de ARI fueron bajos en algunos escenarios, especialmente en remuestreo, lo que revela sensibilidad a variaciones de inicialización y composición. En consecuencia, la estabilidad se interpreta como moderada: suficiente para usar la segmentación como resumen descriptivo institucional, pero insuficiente para afirmar que existe una partición única e inmutable de la población hospitalaria.

Como complemento, se evaluó la reproducibilidad de las métricas internas en las mismas corridas. Las siluetas medias se mantuvieron en rangos moderados: entre 0,251 y 0,254 en no transmisible y entre 0,273 y 0,275 en infecciosa no COVID, según tipo de validación. Esta estabilidad de las métricas internas refuerza que la variabilidad principal no corresponde a cambios extremos de separación promedio, sino a reasignaciones parciales de etiquetas entre perfiles cercanos. El resumen se presenta en la tabla 90.

7.6.1 Calidad del dato y controles técnicos

El análisis de calidad revisó indicadores anuales, tendencia mensual en 2024, datos faltantes, categorías diagnósticas raras y plausibilidad de combinaciones clínicas. Los indicadores anuales mostraron que 2024 no fue claramente mejor que 2021–2023: las mal definidas aumentaron 0,53 puntos porcentuales y la discrepancia clínica aumentó 0,51 puntos porcentuales, ambas diferencias con $p < 0,05$, aunque de baja magnitud absoluta. El indicador obstétrico disminuyó 1,45 puntos porcentuales, sin significación estadística.

Dentro de 2024, la proporción de mal definidas o inespecíficas presentó tendencia descendente significativa: pendiente de -0,4465 puntos porcentuales por mes, $r = -0,8689$, $p = 0,000244$. En cambio, el indicador de obstétricas sin otra especificación no mostró tendencia lineal: pendiente de -0,0204 puntos porcentuales por mes, $p = 0,7681$. Este último indicador se usa como aproximación a especificidad diagnóstica obstétrica: una menor proporción de registros sin otra especificación sugeriría mayor detalle del diagnóstico consignado, aunque no define por sí mismo

buena o mala calidad clínica. En esta versión no se recalculó el agrupamiento por mes ni se cuantificó el efecto de la tendencia mensual de calidad sobre la estructura de los grupos; por tanto, dicho efecto se mantiene como limitación. Las figuras correspondientes se trasladan al apéndice para no sobrecargar el cuerpo del capítulo.

Tabla 49

Datos faltantes en variables clave de la base final

Variable	n total	Faltantes	Faltantes %
Grupo clínico final	26.266	0	0,00
Macrogrupo clínico	26.266	0	0,00
Mortalidad binaria	26.266	0	0,00
Edad	26.266	0	0,00
Grupo de edad	26.266	0	0,00
Sexo	26.266	0	0,00
Código diagnóstico	26.266	0	0,00
Nombre diagnóstico	26.266	0	0,00
CIE-3	26.266	0	0,00
Servicio	26.266	0	0,00
Unidad médica	26.266	0	0,00
Discrepancia clínica	26.266	0	0,00

Nota. La ausencia de faltantes se refiere a las variables clave procesadas en la base final de agrupamiento.

Los diagnósticos raros fueron numerosos en términos de categorías, pero no dominaron la cohorte. En CIE-3 hubo 1.029 categorías, de las cuales 363 tuvieron frecuencia menor o igual a 2; estas representaron 497 episodios, equivalentes al 1,89 % de la cohorte. En nombre diagnóstico hubo 2.355 categorías, 1.365 con frecuencia menor o igual a 2, que concentraron 1.747 episodios, equivalentes al 6,65 %. Esto justifica el uso de CIE-3 como nivel diagnóstico más estable que el

nombre completo. Para evaluar si las categorías CIE-3 extremadamente raras alteraban la solución, se excluyeron códigos CIE-3 con frecuencia menor o igual a 2 dentro de los dos macrogrupos subdivididos y se recalculó la partición. En no transmisible, la partición de los registros retenidos tuvo $ARI=0,950$ y $NMI=0,922$ frente a la solución final; en infecciosa no COVID, $ARI=0,693$ y $NMI=0,708$. Por tanto, la solución no transmisible fue muy robusta a esta exclusión y la infecciosa mostró sensibilidad moderada, coherente con su mayor heterogeneidad y menor estabilidad general. El detalle se reporta en la tabla 89.

Los controles de plausibilidad identificaron 2 registros materno-obstétricos en sexo masculino (0,0076 %), 6 materno-obstétricos con edad menor de 10 o mayor de 60 años (0,0228 %) y 56 diagnósticos neonatales con edad mayor de un año (0,2132 %). Estas frecuencias son bajas, pero deben conservarse como señales para auditoría de codificación. En variables numéricas, edad no presentó valores atípicos por regla IQR; estancia e ingreso sí presentaron colas largas, pero estas variables no fueron usadas para formar los grupos, por lo que no afectaron directamente la agrupamiento.

7.7 Base médica final, limitaciones y conclusión

La última sección sintetiza la estructura de la base resultante, su uso en análisis posteriores y los límites metodológicos que deben acompañar la interpretación del capítulo.

La base médica final contiene 26.266 registros sin COVID-19 y 26.262 identificadores únicos de episodio. Conserva los identificadores clínicos, macrogrupo reconstruido, grupo clínico final, variables demográficas, diagnóstico de egreso, mortalidad, variables administrativas y variables financieras. Esta estructura permite analizar posteriormente servicio, unidad médica, EPS, estancia e ingreso como características asociadas a los grupos, no como variables que los definieron. La presencia de cuatro identificadores de episodio repetidos se reporta como control de trazabilidad; bajo la lógica institucional descrita, estos registros no se interpretan automáticamente como pacientes independientes repetidos, pero deben revisarse antes de usar la base para análisis

longitudinales, auditoría individual o conteos estrictos fila–episodio.

Para evaluar si los grupos finales aportaban información adicional frente a los macrogrupos originales, se compararon modelos logísticos exploratorios de mortalidad con predictores demográficos, macrogrupo y grupo clínico final. La Tabla 50 muestra que el modelo con edad, sexo y grupo final obtuvo mayor AUC y menor log-loss que el modelo con edad, sexo y macrogrupo. Esta evidencia respalda que la segmentación fina añade capacidad descriptiva para diferenciar perfiles de riesgo, aunque no convierte los grupos en predictores causales ni reemplaza la validación clínica externa.

Tabla 50

Utilidad exploratoria de los grupos finales frente a macrogrupos para explicar mortalidad

Modelo	AUC	Log-loss	Brier	AIC aprox.
Edad + sexo	0,712	0,0848	0,0174	1.344,36
Edad + sexo + macrogrupo	0,731	0,0837	0,0174	1.338,79
Edad + sexo + grupo final	0,746	0,0825	0,0173	1.328,34
Solo macrogrupo	0,592	0,0868	0,0175	1.382,31
Solo grupo final	0,652	0,0860	0,0175	1.377,25

Nota. Comparación exploratoria en conjunto de prueba del 30 %. La mortalidad no se usó para formar los grupos. El menor AIC aproximado y el mayor AUC del modelo con grupo final sugieren que la segmentación agrega información descriptiva frente al macrogrupo, sin convertir el análisis en un modelo causal o predictivo definitivo.

7.7.1 Limitaciones metodológicas

La solución tiene alcance exploratorio e institucional. Las principales limitaciones son:

1. El uso de un único año no prueba representatividad interanual.
2. No se dispone de información documental completa sobre cambios institucionales de codificación o contratación.

3. Aunque la unidad operativa es el episodio consolidado, la base final presentó 26.266 filas y 26.262 identificadores únicos de episodio; por tanto, cuatro identificadores aparecen repetidos y deben auditarse si se requiere unicidad estricta fila-episodio.
4. No existe un identificador longitudinal externo al episodio, documento o historia clínica, por lo cual no se evaluaron trayectorias longitudinales de paciente fuera de la definición operativa de episodio.
5. La reclasificación clínica no cuenta con estándar externo independiente.
6. No fue posible ajustar por comorbilidad porque la base disponible no incluye diagnósticos secundarios ni índices de comorbilidad.
7. El paso de microgrupos a centroides introduce una aproximación; se cuantificó de forma descriptiva mediante distancias al centroide, pero no equivale a una validación individual de cada asignación.
8. Las métricas internas muestran separación moderada y solapamiento clínico.
9. La estabilidad ARI/NMI es moderada y sensible en algunos escenarios.
10. La mortalidad ajustada por edad y sexo es un análisis *post hoc* que no equivale a modelo causal ni predictivo.
11. No se ejecutó sensibilidad estructural completa con COVID-19 porque el paquete avanzado no incluyó un dataframe con los 14 casos COVID-19 y etiquetas comparables.
12. No se evaluó formalmente el impacto de la variación temporal de calidad del dato sobre la composición de los grupos porque la base principal disponible no contenía columna de fecha o mes para esa comparación.

7.7.2 *Conclusión del capítulo*

El capítulo construyó una segmentación clínico-diagnóstica de la cohorte 2024 sin COVID-19, compuesta por 26.266 episodios y 10 grupos clínicos finales. La solución separó los dos macrogrupos de mayor interés analítico en tres subgrupos cada uno, manteniendo los demás macrogrupos como grupos únicos por parsimonia y finalidad descriptiva. Las decisiones metodológicas fueron documentadas mediante reglas clínicas, verificación de asignación completa, codificación categórica, SVD, estrategia MiniBatchKMeans–Ward, selección multicriterio de k , sensibilidad SVD, estabilidad con 30 corridas, reproducibilidad de métricas internas, análisis de agregación por microgrupos, sensibilidad ante diagnósticos raros y caracterización *post hoc*.

Los análisis de mortalidad muestran que el grupo no transmisible de adulto mayor cardiovascular-neurológico (No transmisible-C02) tiene la mayor mortalidad cruda local, aunque parte de esta diferencia se reduce al ajustar por edad y sexo. El grupo infeccioso no COVID masculino de mayor mortalidad ajustada por edad y sexo (Infecciosas No COVID- CO1) conserva una mortalidad ajustada solo por edad y sexo elevada, lo que respalda su priorización para revisión clínica. El grupo no transmisible heterogéneo de alto volumen (No transmisible-C03) no tuvo la mortalidad local más alta, pero aportó el mayor número absoluto de muertes por su volumen. La comparación exploratoria entre modelos con macrogrupo y modelos con grupo final mostró mayor AUC y menor log-loss al incorporar la segmentación final, lo que respalda su utilidad descriptiva adicional. La segmentación permite describir perfiles diferenciados de riesgo, carga absoluta y organización clínica-operativa, siempre dentro de los límites de un análisis exploratorio no supervisado y sin

atribuir causalidad a los grupos encontrados.

8. Análisis financiero por grupo clínico

8.1 Objetivo, alcance y exploración cuantitativa inicial

Este capítulo evalúa si los grupos clínicos contruidos previamente presentan diferencias en el ingreso total del episodio hospitalario después de considerar la estancia. El análisis usa los grupos clínicos como factor explicativo posterior; no los usa para crear la segmentación. Esta separación metodológica evita circularidad, porque el algoritmo de agrupamiento no incorporó ingreso, costo, estancia, mortalidad, EPS, servicio ni unidad médica.

La unidad de análisis corresponde al episodio hospitalario. La base financiera incluyó 26.266 episodios. La variable financiera principal fue **Ingreso Total**; la covariable de ajuste fue **Estancia Facturada**. El capítulo interpreta **Ingreso Total** como ingreso facturado o causado del episodio. No interpreta esta variable como costo, utilidad, margen ni rentabilidad. Esta delimitación mantiene separada la evidencia estadística de cualquier conclusión administrativa o financiera neta.

El análisis inició con una exploración cuantitativa de ingreso y estancia. Esta revisión resulta necesaria porque una transformación de escala debe responder a la forma empírica de la variable, a la presencia de valores extremos y al objetivo del modelo, no a una decisión automática (33, 48, 49). La Tabla 51 resume las dos variables continuas utilizadas en el modelo.

Tabla 51*Descripción inicial de las variables cuantitativas del análisis financiero*

Variable	<i>n</i>	Media	Mediana	DE	Mín.	Máx.
Ingreso Total	26.266	5.611.577	1.734.110	12.695.960	95.550	509.569.600
Est Facturada	26.266	4,04	2,00	4,84	1	101

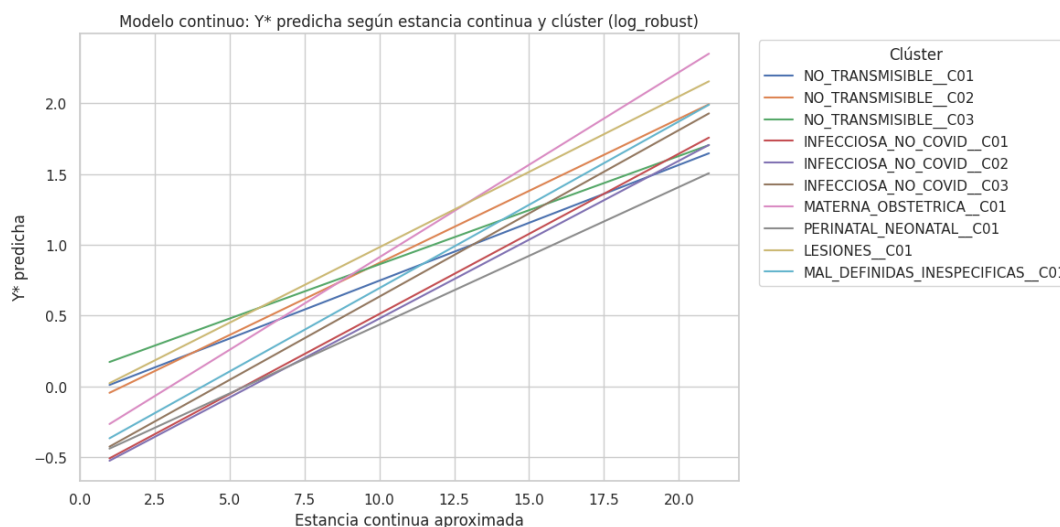
Nota. Los ingresos están expresados en pesos colombianos. La estancia está expresada en días. La diferencia entre media y mediana, junto con los máximos observados, muestra asimetría positiva y presencia de episodios extremos. El análisis complementario incluye histogramas, boxplots, Q-Q plots y diagnósticos de influencia.

La exploración gráfica inicial confirmó que el ingreso total no presenta una distribución simétrica en la escala monetaria original. En el histograma trasladado a los Complementos del Capítulo 8 se observa una alta concentración de episodios en valores bajos y medios, junto con una cola derecha prolongada. Este patrón evidencia que pocos episodios con ingresos muy altos desplazan la distribución y afectan la media; por tanto, no resulta suficiente describir la variable únicamente mediante medidas de tendencia central. En concordancia, la media superó ampliamente la mediana y el máximo observado alcanzó 509.569.600 pesos.

La distribución de $\log(\text{Ingreso Total})$, también presentada en los Complementos del Capítulo 8, muestra que la transformación logarítmica reduce la dominancia visual de los valores extremos y permite observar con mayor claridad el cuerpo central de la distribución. Por esta razón, el modelo principal no utilizó el ingreso original como respuesta directa, sino una respuesta logarítmica robusta, construida a partir del logaritmo del ingreso y posteriormente reescalada mediante mediana e IQR. Los gráficos complementarios de Q-Q plot y boxplot documentan con mayor detalle la forma de la distribución antes y después de la transformación. Además, la estancia se conservó como variable continua porque también presentó cola derecha, con mediana de 2 días y máximo de 101 días; crear rangos de estancia habría reducido información e introducido puntos de corte adicionales.

Figura 2

Relación exploratoria entre estancia y respuesta financiera transformada por grupo clínico.



La Figura 2 funcionó como revisión exploratoria de la relación estancia–respuesta. La nube de puntos mostró una asociación positiva entre estancia e ingreso transformado, aunque no describió una relación perfectamente lineal ni una varianza constante. Por esa razón, el análisis mantuvo la estancia como covariable continua, evaluó explícitamente la interacción grupo clínico por estancia y complementó la inferencia con diagnóstico residual y errores estándar robustos.

8.2 Transformación de la respuesta financiera y preparación de la estancia

El ingreso hospitalario mostró asimetría positiva y valores extremos. En ese contexto, la literatura recomienda considerar transformaciones que estabilicen la escala, reduzcan asimetría y mejoren el comportamiento del modelo de regresión (17, 50, 51). El análisis comparó varias transformaciones candidatas: ingreso original estandarizado, transformación robusta directa, logaritmo, logaritmo robusto, Box–Cox, Yeo–Johnson y winsorización seguida de logaritmo. La comparación no buscó una transformación perfecta; buscó una escala con equilibrio entre forma distribucional, diagnóstico residual, desempeño del modelo e interpretación financiera.

La respuesta principal se definió como una transformación logarítmica con escalamiento

robusto:

$$Y_i^* = \frac{\log(Y_i + c) - \text{Mediana}[\log(Y + c)]}{\text{IQR}[\log(Y + c)]}. \quad (8.1)$$

En esta fórmula, Y_i representa el ingreso total del episodio i . La constante fue $c = 0$ porque la base analizada no contenía ingresos iguales a cero ni ingresos negativos; el ingreso mínimo observado fue 95.550 pesos. Por tanto, el análisis pudo aplicar directamente $\log(Y)$. Si una actualización futura de la base incorpora ingresos nulos o negativos, el equipo deberá revisar esta decisión antes de replicar el modelo.

El modelo seleccionó Y_i^* como escala principal por cuatro razones. Primero, el logaritmo redujo la asimetría del ingreso original. Segundo, el escalamiento por mediana e IQR expresó la respuesta en unidades robustas y redujo la dependencia de la media y la desviación estándar, medidas sensibles a valores extremos (52, 53). Tercero, la escala logarítmica conservó una interpretación financiera relativa más clara que las transformaciones de potencia. Cuarto, el ANCOVA con interacción mantuvo un $R_{\text{aj.}}^2 = 0,481$ bajo la escala logarítmica robusta, mientras que Yeo–Johnson y Box–Cox mejoraron la simetría marginal pero redujeron el $R_{\text{aj.}}^2$ a 0,435. La Tabla 52 resume la comparación principal; el Complemento F.1 presenta la comparación extendida con normalidad residual, Breusch–Pagan y Levene.

Tabla 52*Comparación metodológica de transformaciones de la respuesta financiera*

Transformación	Asim.	Curt.	Asim. resid.	$R^2_{aj.}$	Lectura metodológica
Ingreso original estandarizado	9,458	203,619	10,139	0,355	Conservó asimetría y residuos extremos; no resultó adecuada como escala principal.
Logaritmo	0,692	0,028	0,527	0,481	Redujo asimetría y conservó interpretación financiera relativa.
Logaritmo robusto	0,692	0,028	0,527	0,481	Escala principal; combina logaritmo e interpretación robusta por mediana/IQR.
Yeo–Johnson	0,063	-0,567	-0,152	0,435	Mejóro la simetría marginal, pero redujo el ajuste del ANCOVA.
Box–Cox	0,063	-0,567	-0,152	0,435	Comportamiento equivalente a Yeo–Johnson en esta base.
Winsorización + log	0,660	-0,167	0,496	0,478	Sensibilidad; requiere modificar colas antes del modelo.

Nota. La curtosis corresponde a exceso de curtosis. El diagnóstico extendido por transformación aparece en el Complemento F.1. Ninguna transformación eliminó por completo la heterocedasticidad; por eso el análisis posterior mantuvo diagnóstico residual y errores robustos.

La estancia ingresó al modelo como variable continua estandarizada:

$$\text{estancia}_{z,i} = \frac{\text{estancia}_i - \bar{\text{estancia}}}{s_{\text{estancia}}}. \quad (8.2)$$

Esta estandarización no altera el orden de los episodios ni elimina información. El procedimiento expresa el efecto de la estancia como cambio asociado a una desviación estándar, mejora

la escala numérica del modelo y facilita la comparación de pendientes entre grupos clínicos (54, 55). Para la interpretación clínica, el capítulo también reporta estimaciones marginales ajustadas en días reales: 1 día (P25), 2 días (P50) y 5 días (P75).

8.3 Diseño ANCOVA, supuestos evaluados y diagnóstico del modelo

El análisis usó un modelo ANCOVA porque la pregunta combinó una respuesta continua transformada, un factor categórico de interés y una covariable continua de ajuste. En este caso, el factor fue el grupo clínico y la covariable fue la estancia. Este diseño permite estimar diferencias financieras entre grupos después de considerar la duración de la hospitalización (51, 56).

Antes de adoptar el modelo como herramienta inferencial, el análisis revisó tres condiciones metodológicas. Primero, tomó el episodio hospitalario como unidad de análisis. Por tanto, asumió independencia a nivel de episodio. Si un mismo paciente aportó más de un episodio durante el periodo analizado, esta independencia no equivale a independencia estricta entre individuos. La base utilizada para este capítulo no incorporó un identificador longitudinal de paciente dentro del modelo; por ello, el capítulo reconoce la posible dependencia entre episodios repetidos como una limitación. Segundo, la Figura 2 permitió revisar la relación estancia–respuesta y mostró una asociación positiva compatible con el uso de estancia continua, aunque con dispersión no constante. Tercero, el modelo aditivo actuó como formulación inicial con pendiente común; luego, el modelo con interacción evaluó si esa pendiente común resultaba sostenible.

El modelo aditivo fue:

$$Y_i^* = \beta_0 + \alpha_{\text{grupo}_i} + \beta_1 \text{estancia}_{z,i} + \varepsilon_i. \quad (8.3)$$

El modelo con interacción fue:

$$Y_i^* = \beta_0 + \alpha_{\text{grupo}_i} + \beta_1 \text{estancia}_{z,i} + \gamma_{\text{grupo}_i} \text{estancia}_{z,i} + \varepsilon_i. \quad (8.4)$$

La interacción tuvo justificación clínica y estadística. Desde la perspectiva clínica, una estancia adicional puede implicar intensidades diferentes de procedimientos, medicamentos, dispositivos, soporte crítico o rutas de atención según el perfil clínico. Desde la perspectiva estadística, la interacción contrastó la hipótesis de pendientes homogéneas entre grupos. El resultado global de la interacción rechazó la pendiente común, por lo que el capítulo interpreta el modelo con interacción como especificación principal.

Tabla 53

Comparación de modelos sobre la respuesta logarítmica robusta

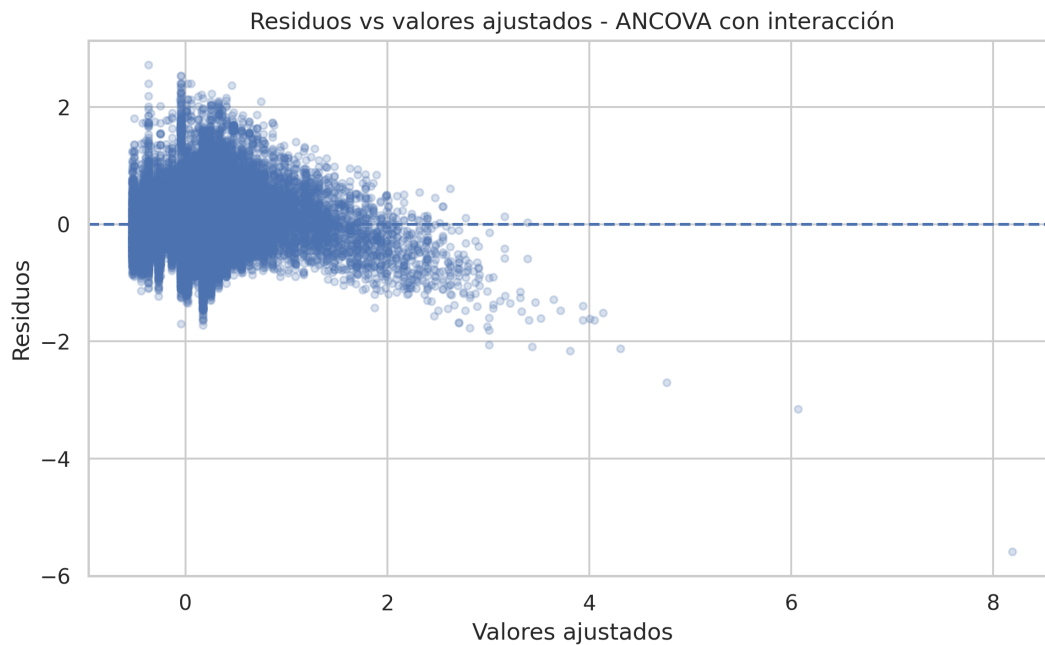
Modelo	R^2	$R^2_{aj.}$	AIC	BIC
ANCOVA con interacción	0,482	0,481	42.052,57	42.216,09
ANCOVA aditiva	0,470	0,470	42.624,00	42.713,94
Estancia sola	0,392	0,392	46.191,34	46.207,69
Grupo clínico solo	0,128	0,127	55.713,22	55.794,98
Modelo nulo	0,000	0,000	59.278,72	59.286,89

Nota. AIC y BIC solo comparan modelos ajustados sobre la misma respuesta transformada; no se usaron para escoger entre escalas diferentes de Y . La selección de escala consideró forma distribucional, diagnóstico residual, desempeño e interpretabilidad (57-59).

Los diagnósticos residuales mostraron que el modelo no cumplió normalidad ni homocedasticidad estrictas. Con 26.266 observaciones, las pruebas formales de normalidad pueden rechazar desviaciones pequeñas; sin embargo, el gráfico de residuos frente a valores ajustados mostró varianza no constante y el Q–Q plot extendido mostró desviaciones en las colas. Además, la distancia de Cook identificó episodios influyentes, incluyendo estancias prolongadas e ingresos altos. El análisis no eliminó esos episodios de forma automática porque pueden representar casos clínico-financieros reales. En cambio, mantuvo el diagnóstico como evidencia de cautela e interpretó las comparaciones posteriores con errores estándar robustos HC3.

Figura 3

Residuos frente a valores ajustados del ANCOVA con interacción.



El análisis escogió HC3 porque esta variante penaliza con mayor fuerza las observaciones de alto apalancamiento que HC0 y HC1, una propiedad útil cuando los diagnósticos muestran heterocedasticidad e influencia (19-21). Las figuras de Q–Q plot, distancia de Cook y tabla de casos influyentes aparecen en el Complemento F.3 para evitar duplicar material gráfico en el cuerpo principal del capítulo.

8.4 Resultados inferenciales y estimaciones marginales ajustadas por estancia real

El modelo con interacción permitió contrastar tres hipótesis globales: diferencias entre grupos ajustadas por estancia, asociación de estancia ajustada por grupo e interacción entre grupo clínico y estancia. La Tabla 54 resume los contrastes.

Tabla 54*Pruebas globales del modelo ANCOVA con interacción*

Contraste	<i>F</i>	gl num.	<i>p</i>	Decisión
Grupo clínico ajustado por estancia	426,662	9	< 0,001	Rechazar H_0
Estancia ajustada por grupo clínico	16.963,385	1	< 0,001	Rechazar H_0
Interacción grupo clínico por estancia	66,182	9	< 0,001	Rechazar H_0

Nota. La respuesta fue Y^* , es decir, ingreso total logarítmico y escalado por mediana/IQR. La significancia de la interacción indica que el efecto de la estancia no mantiene una pendiente común en todos los grupos clínicos.

Las comparaciones post hoc usaron contrastes ajustados del modelo ANCOVA, errores estándar robustos HC3 y corrección de Holm para controlar el error familiar. Holm resulta menos conservador que Bonferroni y mantiene control fuerte del error familiar bajo una estrategia secuencial (40, 60). El análisis no usó Tukey HSD ni Scheffé como estrategia principal porque estos procedimientos se formularon clásicamente para comparaciones de medias en marcos ANOVA con supuestos de homocedasticidad y diseños más balanceados; además, Tukey se orienta a comparaciones por pares de medias y Scheffé protege familias amplias de contrastes, lo que puede resultar innecesariamente conservador cuando los contrastes provienen de un ANCOVA heterocedástico con tamaños desiguales (61-63). Por tanto, el capítulo prefirió contrastes ajustados del modelo con HC3 y Holm.

De las 45 comparaciones posibles entre grupos clínicos, 40 resultaron estadísticamente significativas bajo Holm. Este resultado aporta evidencia de heterogeneidad financiera entre grupos clínicos después de considerar la estancia. La Tabla 55 presenta comparaciones seleccionadas; el complemento reporta la tabla completa.

Tabla 55*Comparaciones post hoc seleccionadas entre grupos clínicos*

Grupo 1	Grupo 2	Mediana 1	Mediana 2	Efecto	Holm <i>p</i>
Infeciosa no COVID C01	No transmisible C03	1.034.519	3.653.059	0,561	< 0,001
Infeciosa no COVID C02	No transmisible C03	1.007.257	3.653.059	0,592	< 0,001
Mal definidas C01	No transmisible C03	1.113.440	3.653.059	0,446	< 0,001
Materna-obstétrica C01	No transmisible C03	1.493.270	3.653.059	0,361	< 0,001
Lesiones C01	No transmisible C03	2.928.863	3.653.059	0,035	0,518

Nota. Las medianas están en pesos. El efecto corresponde a la diferencia ajustada sobre la respuesta transformada.

La tabla completa de 45 comparaciones debe ubicarse en los complementos.

La interacción obliga a interpretar las diferencias financieras en niveles concretos de estancia. Por esta razón, el análisis calculó estimaciones marginales ajustadas en tres niveles reales: 1 día (P25), 2 días (P50) y 5 días (P75). La Tabla 56 resume las estimaciones ajustadas en pesos aproximados. La transformación inversa facilita lectura ejecutiva, pero la inferencia formal pertenece a la escala logarítmica robusta.

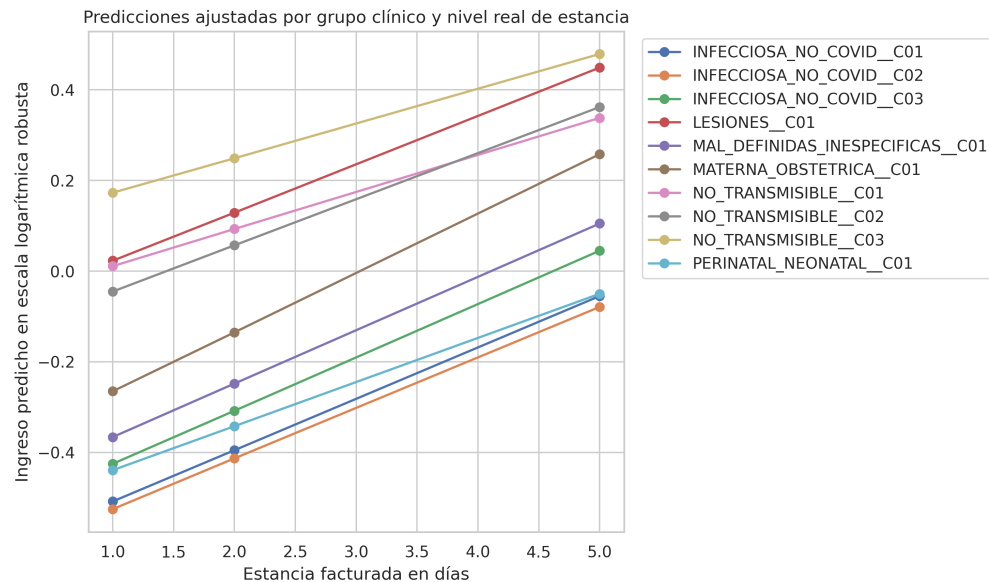
Tabla 56*Estimaciones marginales ajustadas por grupo clínico y nivel real de estancia*

Grupo clínico	1 día	2 días	5 días
Infecciosa no COVID C01	745.705	900.111	1.583.007
Infecciosa no COVID C02	725.342	872.992	1.521.987
Infecciosa no COVID C03	855.574	1.040.325	1.870.268
Lesiones C01	1.801.141	2.150.156	3.657.940
Mal definidas C01	943.600	1.147.588	2.064.338
Materna-obstétrica C01	1.115.743	1.386.562	2.661.116
No transmisible C01	1.765.274	2.022.228	3.040.073
No transmisible C02	1.609.186	1.905.981	3.167.054
No transmisible C03	2.310.497	2.624.042	3.843.855
Perinatal-neonatal C01	836.161	982.787	1.595.763

Nota. Valores aproximados en pesos obtenidos por transformación inversa desde la escala logarítmica robusta. El Complemento F.4 presenta intervalos de confianza para los diez grupos.

Figura 4

Estimaciones marginales ajustadas por grupo clínico y nivel real de estancia.



El patrón completo de la Tabla 56, y no solo los dos grupos de mayor valor, respalda la interpretación de diferencias financieras entre grupos que no dependen únicamente de la duración de la hospitalización. No transmisible C03 presentó las predicciones más altas en los tres escenarios, mientras que Lesiones, No transmisible C02 y No transmisible C01 también conservaron niveles elevados en estancias altas.

8.5 Concentración del ingreso y análisis focalizado del grupo no transmisible C03

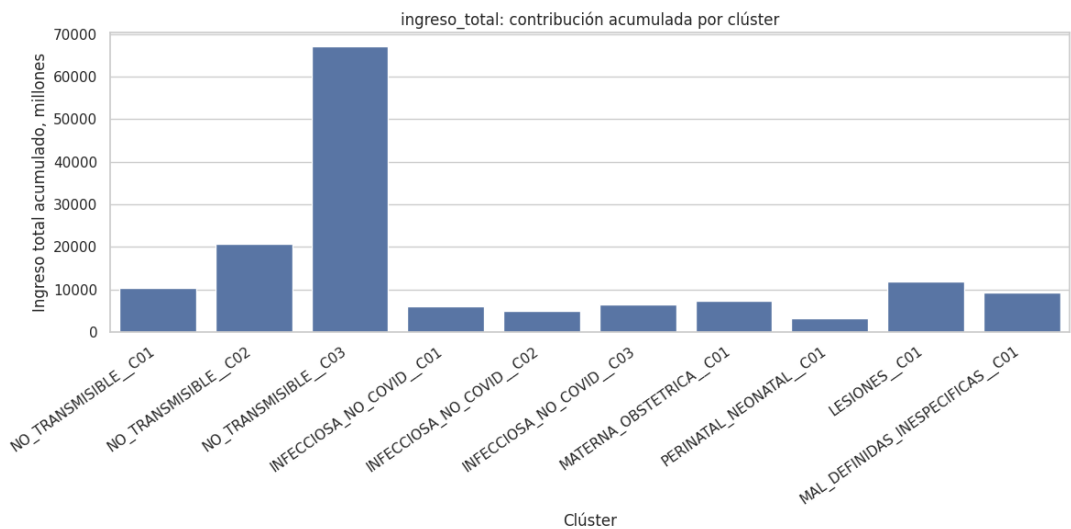
El análisis inferencial respondió si los grupos difieren después de ajustar por estancia. El análisis de contribución acumulada respondió otra pregunta: qué grupos concentran mayor proporción del ingreso total de la cohorte. La Tabla 57 muestra esa concentración.

Tabla 57
Contribución financiera acumulada por grupo clínico

Grupo clínico	<i>n</i>	Ingreso acumulado	Mediana	%
No transmisible C03	7.114	67.061.979.975	3.653.059	45,50
No transmisible C02	2.660	20.649.384.592	2.072.341	14,01
Lesiones C01	1.133	11.835.191.281	2.928.863	8,03
No transmisible C01	1.864	10.409.348.814	2.994.195	7,06
Mal definidas C01	3.328	9.338.316.784	1.113.440	6,34
Materna-obstétrica C01	4.229	7.432.365.272	1.493.270	5,04
Infecciosa no COVID C03	1.823	6.485.232.874	1.411.140	4,40
Infecciosa no COVID C01	1.704	5.975.210.973	1.034.519	4,05
Infecciosa no COVID C02	1.803	4.886.954.704	1.007.257	3,32
Perinatal-neonatal C01	608	3.319.695.491	2.008.966	2,25

Nota. La participación corresponde al porcentaje del ingreso acumulado total explicado por cada grupo.

Figura 5
Participación del ingreso acumulado por grupo clínico.



No transmisible C03 concentró 45,50 % del ingreso acumulado. Este grupo también mantuvo estimaciones marginales ajustadas altas en escenarios de 1, 2 y 5 días. Por tanto, el análisis focalizó este grupo como segmento de mayor peso financiero acumulado.

Dentro de No transmisible C03, los subgrupos con mayor participación fueron neoplasias malignas, enfermedades músculo-esqueléticas, otras neoplasias y enfermedades genitourinarias. La Tabla 58 resume los principales subgrupos; el complemento presenta las figuras y tablas diagnósticas detalladas.

Tabla 58

Subgrupos clínicos con mayor ingreso dentro de No transmisible C03

Subgrupo clínico	<i>n</i>	Ingreso acumulado	Mediana	% ingreso
Neoplasias malignas	1.649	21.711.698.824	7.256.032	32,38
Enfermedades músculo-esqueléticas	998	12.962.516.656	8.916.939	19,33
Otras neoplasias	595	6.856.128.659	5.157.098	10,22
Enfermedades genitourinarias	1.079	5.944.173.083	2.673.490	8,86
Condiciones neuropsiquiátricas	648	4.390.458.539	1.903.273	6,55
Sin clasificación	235	4.353.177.458	4.861.906	6,49
Desórdenes endocrinos	518	3.458.758.212	2.766.193	5,16
Enfermedades respiratorias	556	2.854.372.171	1.572.684	4,26

Nota. La participación se calcula sobre el ingreso acumulado del grupo No transmisible C03.

Este resultado muestra que No transmisible C03 no representa un bloque financiero homogéneo. El grupo combina patologías oncológicas, osteomusculares, renales, respiratorias y neuropsiquiátricas con magnitudes de ingreso diferentes. Por esa razón, el complemento presenta diagnósticos prioritarios por ingreso acumulado, ingreso mediano, variabilidad financiera y

variabilidad de estancia.

8.6 Residuo financiero ajustado y priorización exploratoria

El análisis complementó las comparaciones entre grupos con un modelo de residuo financiero dentro de No transmisible C03. El modelo ajustó el ingreso transformado por diagnóstico y estancia:

$$Y_i^* = \beta_0 + \delta_{\text{diagnóstico}_i} + \beta_1 \text{estancia}_{z,i} + \varepsilon_i. \quad (8.5)$$

Luego calculó el residuo financiero:

$$e_i = Y_{\text{observado},i}^* - \hat{Y}_{\text{esperado},i}^*. \quad (8.6)$$

Un residuo positivo indica que el episodio registró un ingreso superior al esperado en la escala transformada, dado su diagnóstico y estancia. Este residuo no mide rentabilidad, eficiencia ni causalidad. Su uso corresponde a una señal estadística de priorización para revisión posterior.

El análisis evaluó diferencias en el residuo mediante pruebas de Kruskal–Wallis para variables descriptivas. Esta prueba no paramétrica compara distribuciones entre grupos sin exigir normalidad de la respuesta (64, 65). La Tabla 59 resume los resultados.

Tabla 59*Pruebas de Kruskal–Wallis sobre el residuo financiero ajustado*

Variable de grupo	<i>n</i>	Grupos	<i>H</i>	Holm <i>p</i>
Unidad médica	3.201	15	272,346	< 0,001
Clase de movimiento de alta	3.310	3	187,675	< 0,001
EPS	3.148	9	70,987	< 0,001
Servicio	3.333	3	25,304	< 0,001
Grupo de edad	3.336	17	33,328	0,013
Sexo	3.336	2	0,531	0,466

Nota. La respuesta fue el residuo financiero ajustado por diagnóstico y estancia. Estas pruebas cumplen una función exploratoria; no controlan simultáneamente contratos, complejidad, rutas clínicas ni costos reales.

8.6.0.1 Sensibilidad de la regla exploratoria P75/P90.. Los puntos de corte P75 y P90 no corresponden a umbrales clínicos validados ni a estándares externos de auditoría financiera. El análisis los utilizó como una regla operativa inicial de priorización: P75 identifica el cuarto superior de episodios con mayor residuo financiero ajustado, mientras que P90 restringe la categoría de exceso muy alto al decil superior. Para evaluar la dependencia de esta regla frente a la elección de percentiles, se comparó P75/P90 con cuatro escenarios alternativos: P70/P90, P80/P90, P75/P95 y P80/P95.

El análisis de sensibilidad mostró estabilidad interna moderada-alta. La categoría de exceso muy alto mantuvo 334 episodios cuando se conservó P90 como punto de corte superior, y se redujo a 167 episodios al usar P95. En este último caso, el ingreso mediano aumentó de 15.303.739 a 18.981.141 pesos, lo cual es consistente con una regla de priorización más restrictiva. Además, la concordancia entre P75/P90 y los escenarios alternativos presentó valores de ARI entre 0,787 y 0,958, y valores de NMI entre 0,668 y 0,872.

Estos resultados respaldan la estabilidad interna de la clasificación ante cambios razonables

de percentil. Sin embargo, no constituyen validación externa. Por tanto, el capítulo interpreta la clasificación como una herramienta exploratoria para ordenar episodios según residuo financiero ajustado, no como una escala clínica, contractual o financiera validada. El detalle completo del análisis de sensibilidad se presenta en los Complementos del Capítulo 8, Sección 98.

Tabla 60

Priorización exploratoria de episodios según residuo financiero positivo ajustado

Clasificación	<i>n</i>	Ingreso mediano	Estancia mediana	Residuo mediano
Residuo financiero positivo muy alto	334	15.303.739	3,0	0,654
Residuo financiero positivo alto	500	10.083.253	3,0	0,369
Dentro de lo esperado o bajo	2.502	2.173.351	2,0	-0,098

Nota. La clasificación deriva de percentiles del residuo financiero ajustado. El Complemento F.7 presenta sensibilidad de umbrales y diagnósticos frecuentes en exceso muy alto.

La estabilidad interna fue moderada-alta frente a umbrales alternativos: el ARI frente a P75/P90 fue 0,827 con P70/P90, 0,820 con P80/P90, 0,958 con P75/P95 y 0,787 con P80/P95. Esta evidencia respalda estabilidad interna, pero no sustituye validación externa. Para validar utilidad operativa, la institución tendría que contrastar estos episodios con auditorías reales, costos confirmados, glosas, uso de insumos, procedimientos y contratos.

8.7 Síntesis, limitaciones e implicaciones de gestión

El análisis aporta evidencia de diferencias financieras entre grupos clínicos después de considerar la estancia. El modelo ANCOVA con interacción alcanzó $R_{aj.}^2 = 0,481$, superó al modelo aditivo y rechazó la hipótesis de pendientes homogéneas entre grupos. Las estimaciones marginales ajustadas en 1, 2 y 5 días mostraron que No transmisible C03 mantuvo los mayores

ingresos predichos, mientras que Lesiones y otros grupos no transmisibles también conservaron valores altos en escenarios de mayor estancia.

El análisis de contribución acumulada identificó a No transmisible C03 como el grupo de mayor peso financiero, con 45,50 % del ingreso total. Dentro de este grupo, las neoplasias malignas, enfermedades músculo-esqueléticas, otras neoplasias y enfermedades genitourinarias explicaron una proporción importante del ingreso acumulado. Esta composición interna sugiere que el peso financiero del grupo no depende solo de la estancia; también refleja mezcla diagnóstica, complejidad y posibles diferencias en procedimientos o tecnologías.

El capítulo mantiene cuatro limitaciones centrales. Primero, la unidad fue el episodio hospitalario; la ausencia de un identificador longitudinal en el modelo impidió verificar independencia estricta entre pacientes. Segundo, los diagnósticos residuales mostraron heterocedasticidad, desviaciones en colas y observaciones influyentes; por eso el análisis usó HC3 y reportó diagnósticos, pero no eliminó casos reales extremos. Tercero, el ingreso total no mide costos ni rentabilidad. Cuarto, la clasificación P75/P90 del residuo financiero tiene estabilidad interna, pero todavía no cuenta con validación externa institucional.

9. Conclusiones generales, limitaciones y recomendaciones

9.1 Conclusiones alineadas con los objetivos específicos

Objetivo específico 1: caracterización de la población hospitalaria

Se cumplió el objetivo de caracterizar la población de episodios hospitalarios mediante variables clínicas, operativas y financieras. El análisis permitió documentar la estructura de la base 2021–2024, la cohorte principal 2024, la distribución de variables demográficas, diagnósticas, administrativas y financieras, la presencia de alta cardinalidad diagnóstica, la necesidad de recodificaciones trazables y la delimitación del episodio como unidad disponible de análisis. La

evidencia correspondiente se presenta principalmente en los Capítulos 5 y 7.

Objetivo específico 2: análisis exploratorio del ingreso por episodio

Se cumplió el objetivo de analizar exploratoriamente el ingreso por episodio según variables demográficas, diagnósticas, administrativas y operativas. El análisis evidenció asimetría marcada del ingreso, presencia de valores extremos, relación positiva entre estancia e ingreso transformado y diferencias descriptivas entre grupos clínicos. También permitió justificar la transformación logarítmica robusta y el uso de la estancia como covariable continua. La evidencia principal se presenta en los Capítulos 5 y 8.

Objetivo específico 3: segmentación diagnóstica mediante agrupamiento no supervisado

Se cumplió el objetivo de segmentar episodios hospitalarios según variables diagnósticas y demográficas mediante agrupamiento no supervisado. La solución final organizó la cohorte 2024 sin COVID-19 en diez grupos clínicos: tres no transmisibles, tres infecciosos no COVID y cuatro macrogrupos conservados como grupos únicos. La solución fue construida sin utilizar mortalidad, ingreso, estancia, EPS, servicio ni unidad médica, lo que permitió interpretar estas variables posteriormente sin circularidad metodológica. La validación final mostró estabilidad promedio moderada, con sensibilidad en escenarios mínimos, especialmente en el macrogrupo infeccioso. Por tanto, los grupos deben interpretarse como perfiles clínico-diagnósticos institucionales y no como clases clínicas definitivas.

Objetivo específico 4: caracterización financiera de los segmentos

Se cumplió el objetivo de caracterizar financieramente los segmentos identificados. El análisis mostró diferencias de ingreso entre grupos clínicos después de ajustar por estancia, así

como una participación acumulada relevante de los grupos no transmisibles en el ingreso total. En particular, No transmisible-C03 concentró el mayor peso financiero acumulado, mientras que No transmisible-C02 destacó por mortalidad local elevada. Estos resultados permiten priorizar revisiones internas, pero no deben interpretarse como evidencia de rentabilidad, dado que el estudio trabajó con ingreso y no con costos ni utilidad.

Objetivo específico 5: modelos explicativos ajustados del ingreso

Se cumplió el objetivo de desarrollar y evaluar modelos explicativos ajustados del ingreso por episodio. El análisis incluyó ANCOVA con estancia continua, interacción grupo clínico por estancia, errores estándar robustos HC3, corrección de Holm, medias ajustadas y residuos financieros ajustados por diagnóstico y estancia. Estos modelos apoyan la interpretación financiera y la priorización de episodios con ingreso observado superior al esperado, pero no constituyen un sistema predictivo supervisado ni una herramienta de pronóstico individual.

9.2 Limitaciones del estudio

1. El análisis se basó en registros administrativos y clínicos secundarios; por tanto, depende de la calidad de codificación, completitud y consistencia del dato institucional.
2. La unidad disponible fue el episodio hospitalario. No se contó con identificador longitudinal de paciente para evaluar dependencia entre episodios de un mismo individuo, reingresos independientes o trayectorias de seguimiento.
3. La cohorte principal se restringió a 2024 por actualidad institucional y recomendación clínica-operativa. Los resultados no deben generalizarse automáticamente a 2021–2023 ni a años posteriores sin validación temporal.
4. La reclasificación clínica fue reproducible y basada en CIE-10/texto diagnóstico, pero no

reemplaza una auditoría clínica individual ni una validación experta formal caso a caso.

5. La estabilidad del agrupamiento fue promedio moderada y presentó sensibilidad en escenarios mínimos; por tanto, la segmentación debe usarse como herramienta descriptiva institucional, no como clasificación inmutable.
6. La mortalidad ajustada solo controló edad y sexo; no incorporó comorbilidades, severidad, diagnósticos secundarios, procedimientos ni complejidad clínica completa.
7. El análisis financiero utilizó ingreso total, no costos ni utilidad. En consecuencia, no permite concluir rentabilidad, margen financiero ni eficiencia económica neta.
8. Los residuos financieros positivos son señales exploratorias de priorización; no prueban cobro indebido, sobreutilización, causalidad, error administrativo ni rentabilidad.

9.3 Recomendaciones

A partir de los resultados obtenidos, se proponen las siguientes recomendaciones para fortalecer el uso institucional de la segmentación clínico-financiera:

1. Conservar la separación metodológica entre variables usadas para construir los grupos y variables utilizadas para caracterizarlos posteriormente, especialmente ingreso, estancia, mortalidad, EPS, servicio y unidad médica.
2. Usar los grupos clínicos finales como herramienta de priorización descriptiva y no como clasificación diagnóstica definitiva, modelo causal o mecanismo automático de decisión administrativa.
3. Priorizar revisiones clínicas y operativas en los grupos con mayor mortalidad ajustada, mayor aporte absoluto a fallecimientos o mayor variabilidad financiera, en particular Infecciosas C01, No transmisible C02 y No trasmisible C03.

4. Complementar en futuras etapas el análisis de ingreso con información detallada de costos, con el fin de avanzar hacia métricas de margen, rentabilidad o eficiencia económica neta.
5. Fortalecer la calidad del dato diagnóstico mediante auditoría de categorías mal definidas, discrepancias entre clasificación original y reconstruida, y registros con combinaciones clínicas poco plausibles.
6. Reproducir la segmentación en nuevos cortes temporales para evaluar estabilidad interanual, cambios en la mezcla clínica y posibles efectos de modificaciones institucionales en codificación, contratación o rutas de atención.
7. Incorporar diagnósticos secundarios, comorbilidades e identificadores longitudinales de paciente si la institución desea pasar de una lectura por episodio a una evaluación de trayectorias clínicas y riesgo individual.

Estas recomendaciones deben entenderse como líneas de continuidad del trabajo y no como decisiones operativas automáticas. Su aplicación requiere revisión conjunta con los equipos clínicos, financieros, administrativos y de inteligencia empresarial de la institución.

Referencias Bibliográficas

1. Rodríguez KS. Formato protocolo proyecto Comité de Ética: Segmentación de diagnósticos y análisis financiero de pacientes en el servicio de hospitalización mediante modelos de clustering y predictivos de ingreso por paciente. Documento institucional. Presentado al Comité de Ética; director de práctica: Andrés Ríos; codirector de práctica: Jorge Villamizar; asesor médico: Julián Navas. Clínica FOSCAL, 2024
2. Codd EF. A relational model of data for large shared data banks. *Communications of the ACM* 1970; 13:377-87. doi: 10.1145/362384.362685
3. Han J, Kamber M y Pei J. *Data mining: Concepts and techniques*. 3.^a ed. Morgan Kaufmann, 2011
4. Kaufman L y Rousseeuw PJ. *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, 1990
5. Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 2010; 31:651-66. doi: 10.1016/j.patrec.2009.09.011
6. Deerwester S, Dumais ST, Furnas GW, Landauer TK y Harshman R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 1990; 41:391-407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
7. Halko N, Martinsson PG y Tropp JA. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* 2011; 53:217-88. doi: 10.1137/090771806
8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011; 12:2825-30

9. Sculley D. Web-scale k-means clustering. *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010:1177-8. DOI: 10.1145/1772690.1772862
10. Ward Jr JH. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 1963; 58:236-44. DOI: 10.1080/01621459.1963.10500845
11. Murtagh F y Legendre P. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification* 2014; 31:274-95. DOI: 10.1007/s00357-014-9161-z
12. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987; 20:53-65. DOI: 10.1016/0377-0427(87)90125-7
13. Caliński T y Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics* 1974; 3:1-27. DOI: 10.1080/03610927408827101
14. Davies DL y Bouldin DW. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1979; PAMI-1:224-7. DOI: 10.1109/TPAMI.1979.4766909
15. Hubert L y Arabie P. Comparing partitions. *Journal of Classification* 1985; 2:193-218. DOI: 10.1007/BF01908075
16. Vinh NX, Epps J y Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* 2010; 11:2837-54
17. Box GEP y Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 1964; 26:211-52
18. Yeo IK y Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika* 2000; 87:954-9. DOI: 10.1093/biomet/87.4.954

19. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; 48:817-38. DOI: 10.2307/1912934
20. MacKinnon JG y White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* 1985; 29:305-25. DOI: 10.1016/0304-4076(85)90158-7
21. Long JS y Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician* 2000; 54:217-24. DOI: 10.1080/00031305.2000.10474549
22. Lorenzoni L y Marino A. Understanding variations in hospital length of stay and cost: Results of a pilot project. *OECD Health Working Papers* 94. OECD Publishing, 2017. DOI: 10.1787/ae3a5ce9-en
23. Kaplan RS y Porter ME. How to solve the cost crisis in health care. *Harvard Business Review* 2011; 89:46-52
24. Porter ME. What is value in health care? *New England Journal of Medicine* 2010; 363:2477-81. DOI: 10.1056/NEJMp1011024
25. Khalid S, Yang P y Black W. Unsupervised machine learning for data-driven stratification of health care populations: A systematic review. *Health Services Research* 2020
26. Eckhardt L, Snider A y Wensley S. Unsupervised learning techniques for biomedical data analysis: A systematic review. *Bioinformatics* 2020
27. Yan S, Kwan YH, Tan CS, Thumboo J y Low LL. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Medical Research Methodology* 2018; 18:121. DOI: 10.1186/s12874-018-0584-9
28. República de Colombia. Ley Estatutaria 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales [Internet]. *Diario Oficial* No. 48.587. 2012 oct 17. Disponible en: <https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/1684507> [citado 21 ago 2026]

29. República de Colombia. Decreto 1377 de 2013. Por el cual se reglamenta parcialmente la Ley 1581 de 2012 [Internet]. Diario Oficial No. 48.834. 2013. Disponible en: <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=53646> [citado 21 ago 2026]
30. República de Colombia. Ministerio de Salud. Resolución 1995 de 1999. Por la cual se establecen normas para el manejo de la historia clínica [Internet]. Diario Oficial No. 43.655. 1999 jul 8. Disponible en: https://www.minsalud.gov.co/Normatividad_Nuevo/Resoluci%C3%B3n%201995%20de%201999.pdf [citado 21 ago 2026]
31. Vandenbroucke JP, Elm E von, Altman DG, Gøtzsche PC, Mulrow CD et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. PLoS Medicine 2007; 4:e297. DOI: 10.1371/journal.pmed.0040297
32. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Medicine 2015; 12:e1001885. DOI: 10.1371/journal.pmed.1001885
33. Tukey JW. Exploratory data analysis. Addison-Wesley, 1977
34. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 1900; 50:157-75. DOI: 10.1080/14786440009463897
35. Agresti A. Categorical data analysis. 3.^a ed. John Wiley & Sons, 2013
36. Cramér H. Mathematical methods of statistics. Princeton University Press, 1946
37. Benjamini Y y Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 1995; 57:289-300. DOI: 10.1111/j.2517-6161.1995.tb02031.x
38. Efron B y Tibshirani RJ. An introduction to the bootstrap. Chapman & Hall/CRC, 1993

39. World Health Organization. International statistical classification of diseases and related health problems. 10th rev. World Health Organization, 2019
40. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; 6:65-70
41. McKinney W. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*. 2010:56-61. DOI: 10.25080/Majora-92bf1922-00a
42. Harris CR, Millman KJ, Walt SJ van der, Gommers R, Virtanen P et al. Array programming with NumPy. *Nature* 2020; 585:357-62. DOI: 10.1038/s41586-020-2649-2
43. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 2020; 17:261-72. DOI: 10.1038/s41592-019-0686-2
44. Seabold S y Perktold J. Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*. 2010:92-6
45. Hunter JD. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 2007; 9:90-5. DOI: 10.1109/MCSE.2007.55
46. Waskom ML. Seaborn: Statistical data visualization. *Journal of Open Source Software* 2021; 6:3021. DOI: 10.21105/joss.03021
47. Friendly M. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association* 1994; 89:190-200. DOI: 10.1080/01621459.1994.10476460
48. Cleveland WS. Visualizing data. Hobart Press, 1993
49. Fox J y Weisberg S. An R companion to applied regression. 3.^a ed. Sage, 2019
50. Carroll RJ y Ruppert D. Transformation and weighting in regression. Chapman & Hall, 1988
51. Fox J. Applied regression analysis and generalized linear models. 3.^a ed. Sage, 2016
52. Huber PJ y Ronchetti EM. Robust statistics. 2.^a ed. John Wiley & Sons, 2009

53. Wilcoxon RR. Introduction to robust estimation and hypothesis testing. 4.^a ed. Academic Press, 2017
54. Gelman A. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 2008; 27:2865-73. DOI: 10.1002/sim.3107
55. Gelman A e Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, 2007
56. Kutner MH, Nachtsheim CJ, Neter J y Li W. Applied linear statistical models. 5.^a ed. McGraw-Hill/Irwin, 2005
57. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; 19:716-23. DOI: 10.1109/TAC.1974.1100705
58. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; 6:461-4. DOI: 10.1214/aos/1176344136
59. Burnham KP y Anderson DR. Model selection and multimodel inference: A practical information-theoretic approach. 2.^a ed. Springer, 2002
60. Shaffer JP. Multiple hypothesis testing. *Annual Review of Psychology* 1995; 46:561-84. DOI: 10.1146/annurev.ps.46.020195.003021
61. Tukey JW. Comparing individual means in the analysis of variance. *Biometrics* 1949; 5:99-114. DOI: 10.2307/3001913
62. Scheffé H. A method for judging all contrasts in the analysis of variance. *Biometrika* 1953; 40:87-104. DOI: 10.1093/biomet/40.1-2.87
63. Hsu JC. Multiple comparisons: Theory and methods. Chapman & Hall/CRC, 1996
64. Kruskal WH y Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952; 47:583-621. DOI: 10.1080/01621459.1952.10483441
65. Conover WJ. Practical nonparametric statistics. 3.^a ed. John Wiley & Sons, 1999

Apéndices

Apéndice A. Aprobación del Comité de Ética de FOSCAL

En este apéndice se presenta el soporte institucional de aprobación emitido por el Comité de Ética de FOSCAL para el desarrollo del proyecto. Este documento respalda el uso de la información anonimizada suministrada por la institución y la ejecución del análisis bajo los lineamientos éticos correspondientes.

Figura 6

Aprobación del Comité de Ética de FOSCAL para el proyecto.

Protocolo / Universidades					Estado: Decisión emitida			
No. de identificación					Tipo de protocolo			
CEI-FOS.010147.001					Protocolo Inicial			
Título					Patrocinador			
000 Segmentación de diagnósticos y análisis financiero de pacientes en el servicio de hospitalización mediante modelos de clustering y predictivos de ingreso por paciente.					No aplica			
Investigadores					Protocolo			
Keren Slendy Rodriguez Calderón; Andres Sebastián Rios Gutiérrez; Jorge Villamizar Morales; Julian Alonso Navas					010147			
Institución					País		Resultado	
Universidad Industrial de Santander-UIS					COLOMBIA		Aprobado	
Presentado en	Actualizado en	Fecha de Sesión	Fecha de generación del dictamen	Informado en	Fecha de reclutamiento	Finalizado	Monitoreo	
09/02/2026	25/02/2026	16/02/2026	25/02/2026	25/02/2026	No ha iniciado	25/02/2026	No aplica	

Nota. Tomado de: Comité de Ética de FOSCAL (1).

La aprobación del Comité de Ética constituye un soporte formal para la realización del estudio con bases de datos institucionales anonimizadas. En concordancia con el alcance metodológico del proyecto, no se realizó contacto directo con pacientes, no se intervino la atención clínica y los resultados fueron presentados de manera agregada, sin identificación individual.

Apéndice B. Trazabilidad administrativa complementaria

La selección de la cohorte principal 2024 se sustentó en la actualidad institucional de la información y en la recomendación técnica realizada en la reunión del 6 de abril de 2026 a las 11:00 a.m. por el médico Julián Navas, asesor médico del proyecto. Esta trazabilidad se conserva en

los apéndices porque funciona como respaldo administrativo de la decisión, pero no se interpreta como validación clínica externa formal ni como prueba de representatividad estadística del año 2024 frente al periodo 2021–2024.

Apéndice C. Complementos del Capítulo 5

C.1 Figuras descriptivas y visualizaciones complementarias del análisis preliminar

Este complemento reúne las figuras descriptivas, exploratorias y de apoyo visual asociadas al Capítulo 5. Las figuras se trasladan fuera del cuerpo principal para mejorar el flujo narrativo del capítulo, sin eliminar información ni modificar etiquetas, títulos o referencias.

C.1.1 Integridad, faltantes y cardinalidad

Figura 7

Porcentaje de valores faltantes por variable en la base analítica 2021–2024.

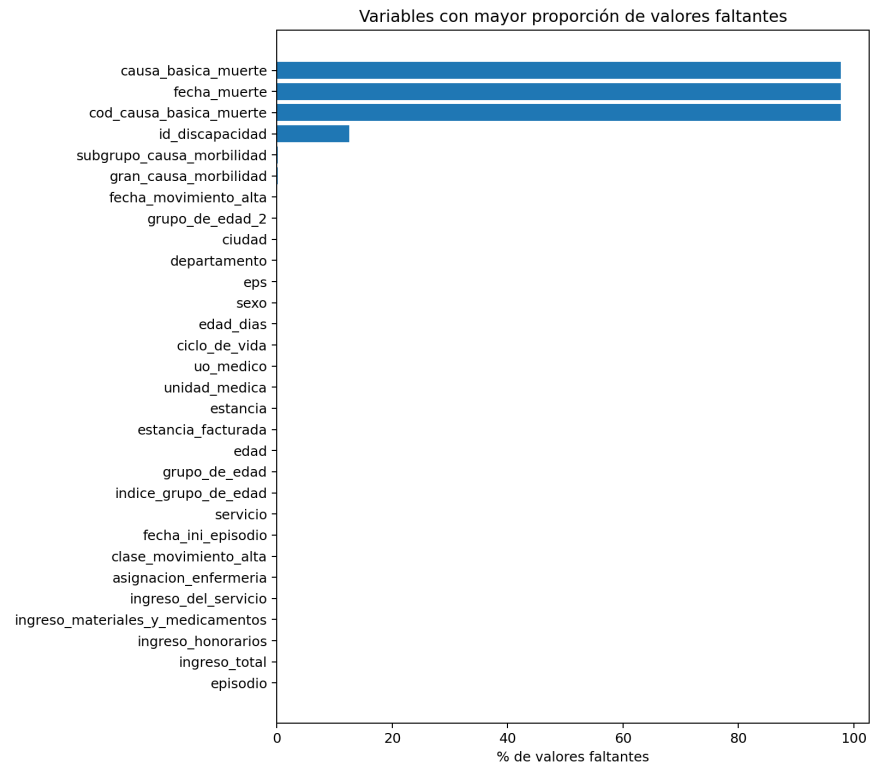
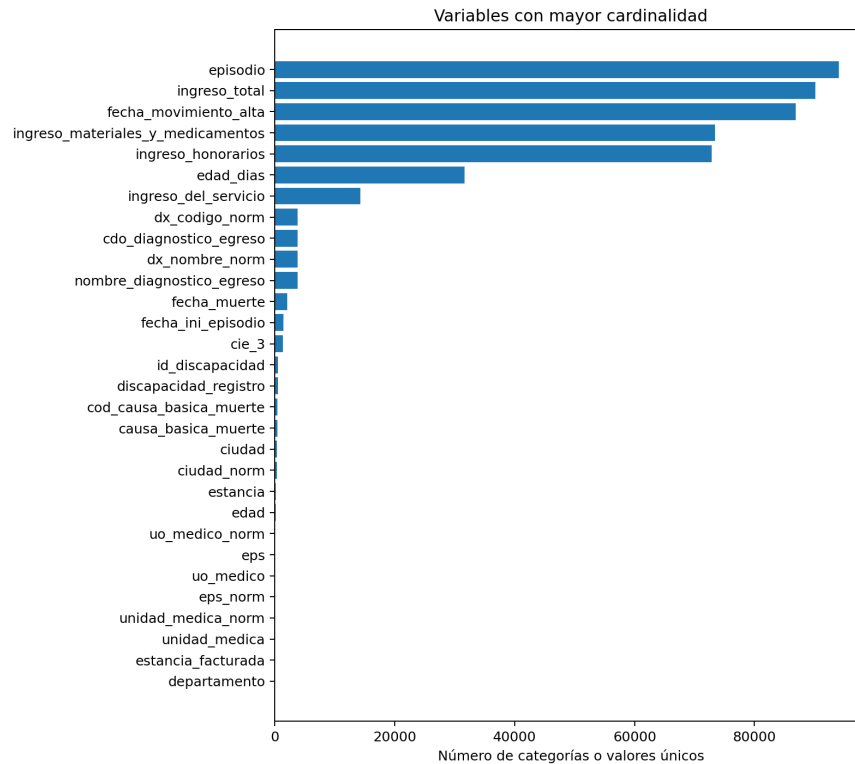


Figura 8

Cardinalidad de las variables disponibles para el análisis preliminar.



C.1.2 Distribuciones preliminares y concentración diagnóstica

Figura 9

Distribución de la edad de los episodios hospitalarios.

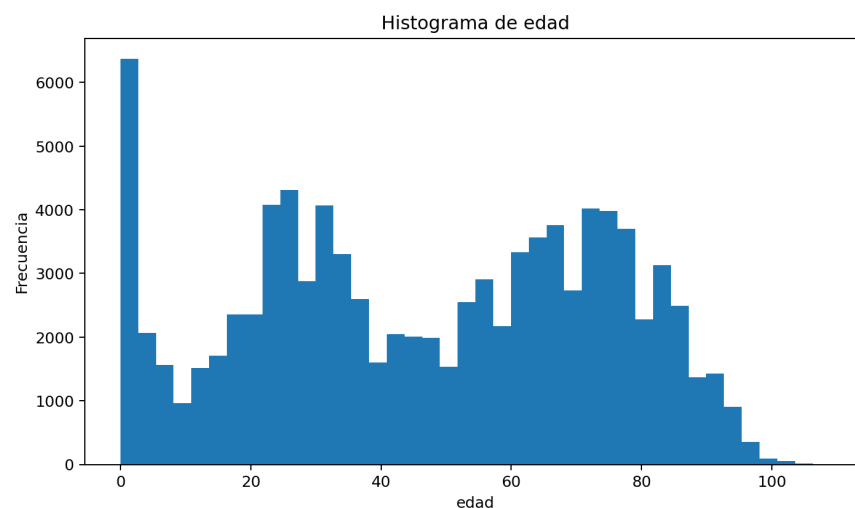
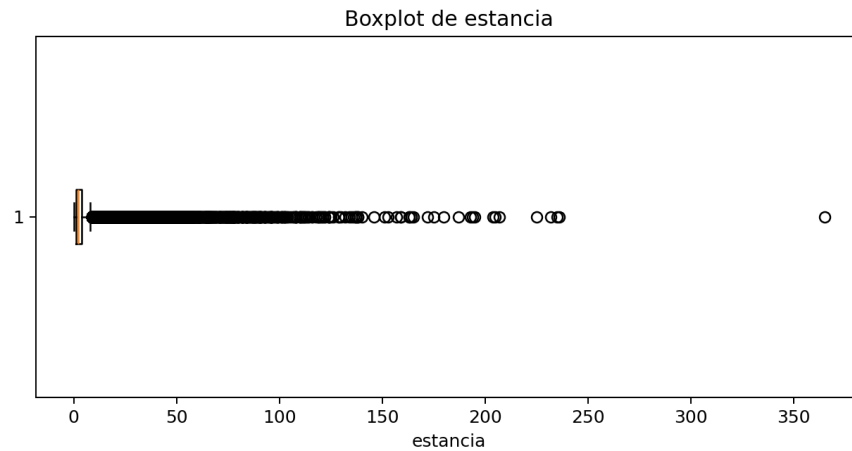
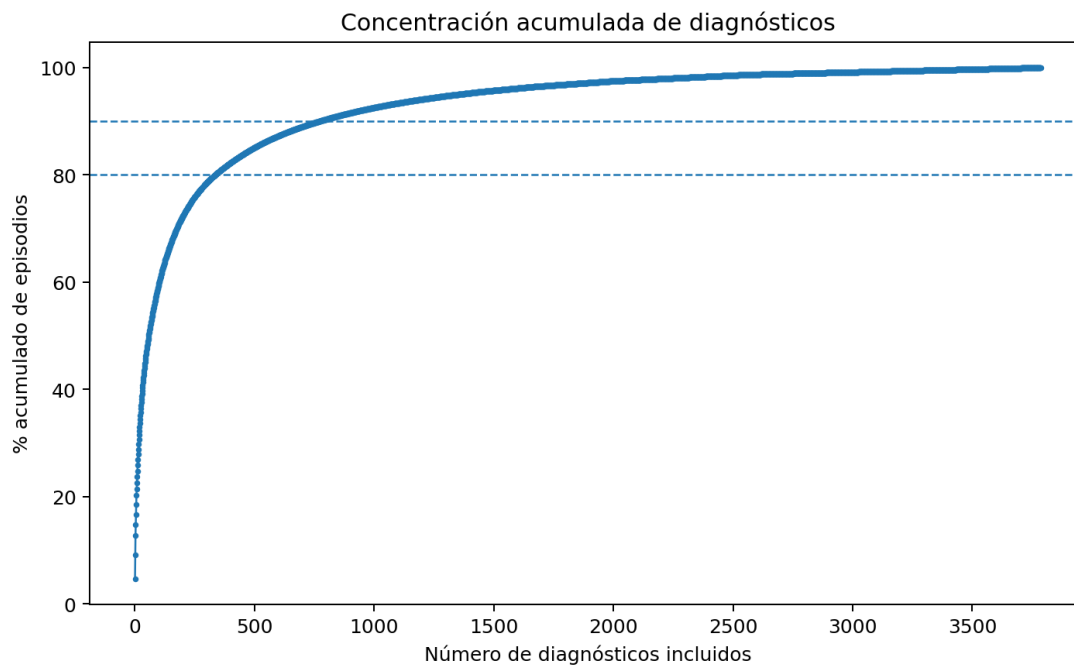


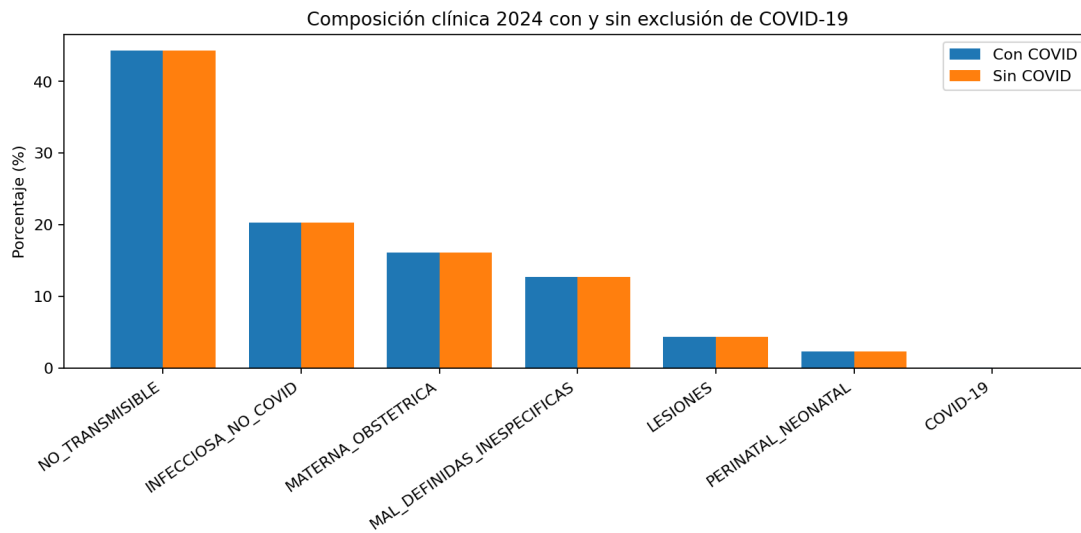
Figura 10*Distribución de la estancia hospitalaria y presencia de valores extremos.***Figura 11***Concentración acumulada de los diagnósticos de egreso.*

La concentración diagnóstica mostró una estructura de alta cardinalidad: un número reducido de diagnósticos agrupó una proporción importante de episodios, mientras que muchas categorías presentaron baja frecuencia. Este comportamiento sustentó la construcción de variables diagnósticas Top- N y la evaluación de sensibilidad frente a distintos puntos de corte.

C.1.3 Sensibilidad de exclusión COVID-19

Figura 12

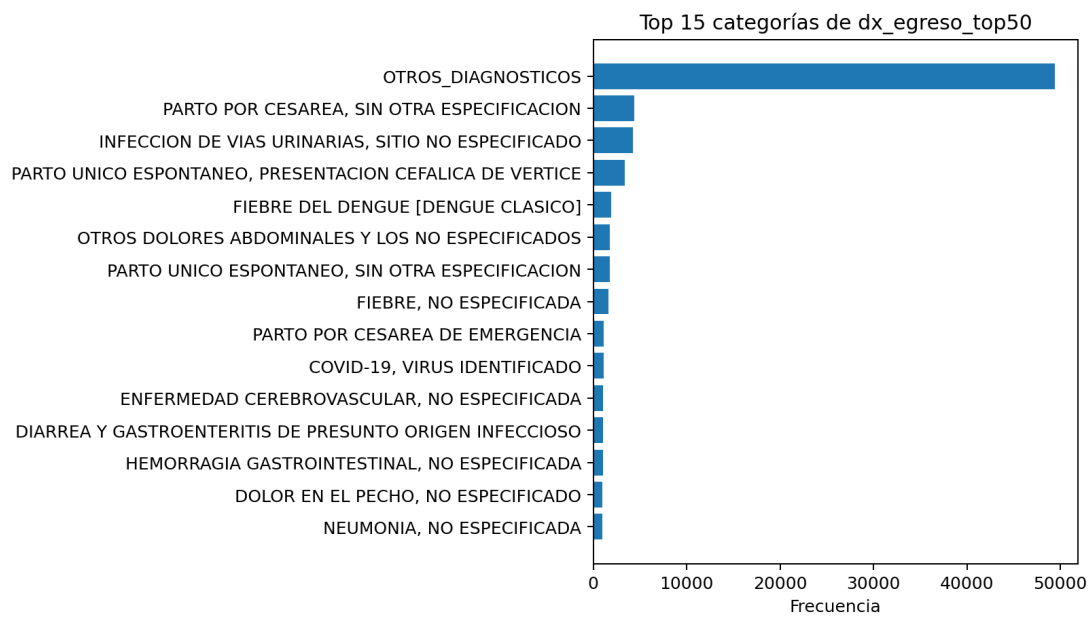
Sensibilidad de la composición de macrogrupos clínicos con y sin episodios COVID-19.



C.1.4 Recodificación diagnóstica

Figura 13

Distribución de las categorías principales del diagnóstico de egreso después de la recodificación Top 50.



C.1.5 Variables clínicas derivadas para agrupamiento

Figura 14

Distribución de la variable grupo de edad utilizada en el proceso de agrupamiento.

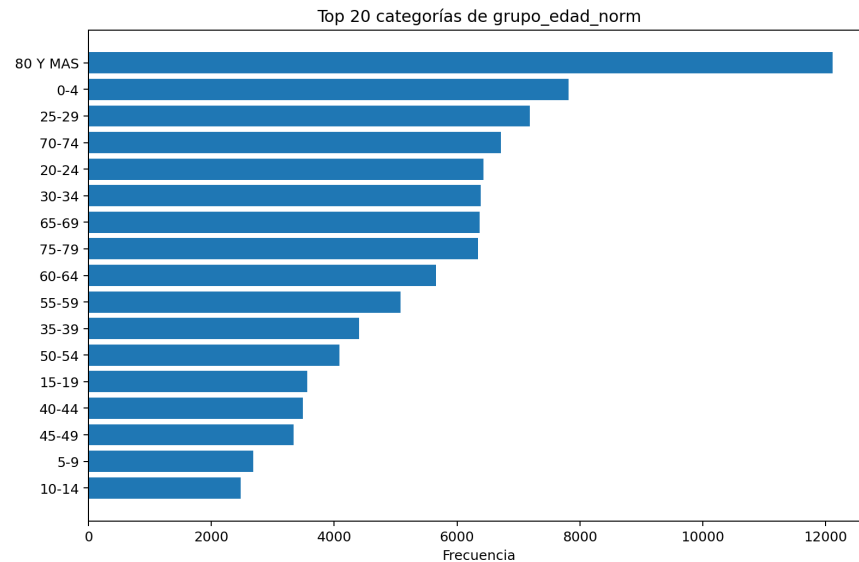
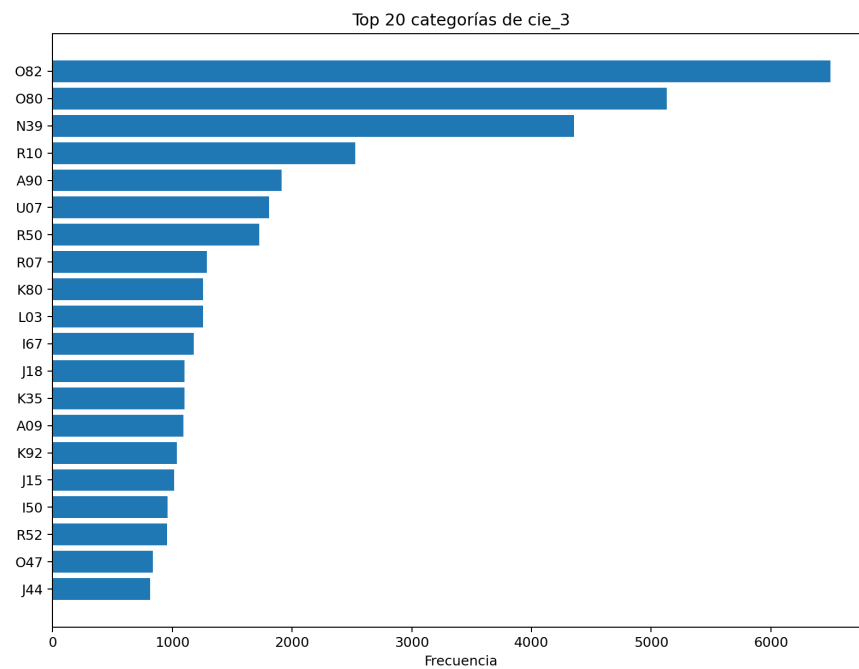


Figura 15

Distribución de los principales prefijos CIE-3 utilizados en el proceso de agrupamiento.



C.1.6 Sensibilidad de reducción dimensional SVD

Figura 16

Curva de varianza explicada acumulada por número de componentes SVD en el macrogrupo no transmisible.

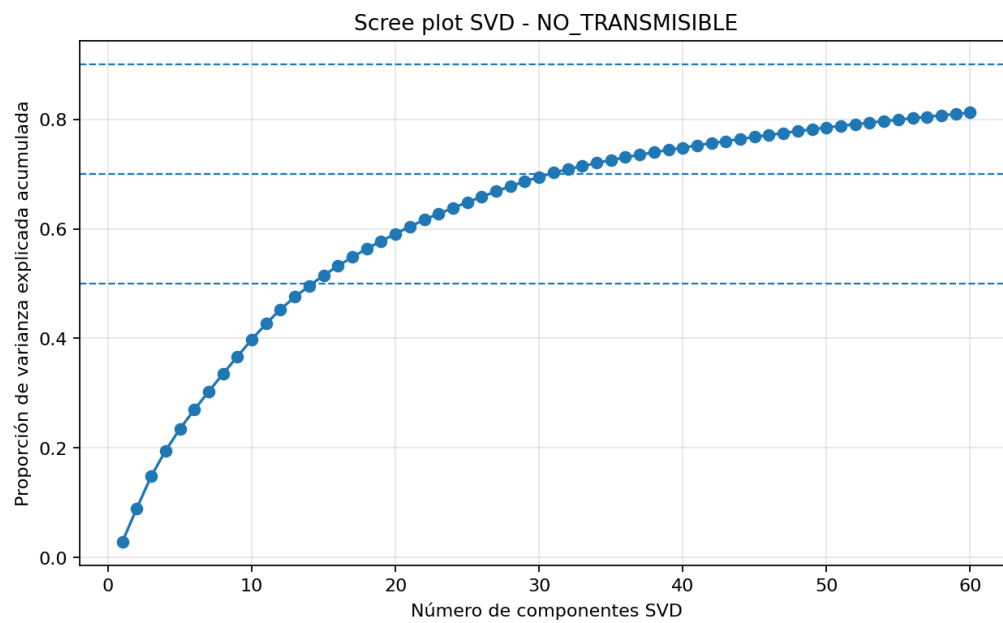
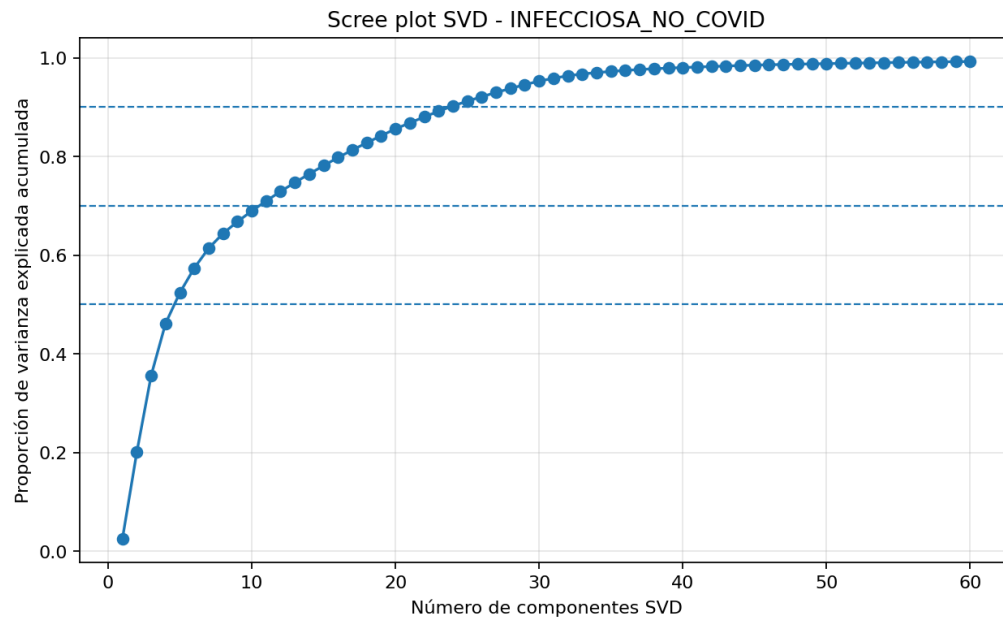


Figura 17

Curva de varianza explicada acumulada por número de componentes SVD en el macrogrupo infecciosa no COVID.



C.1.7 Selección del número de grupos

Figura 18

Coeficiente de silueta según el número de grupos en el macrogrupo no transmisible.

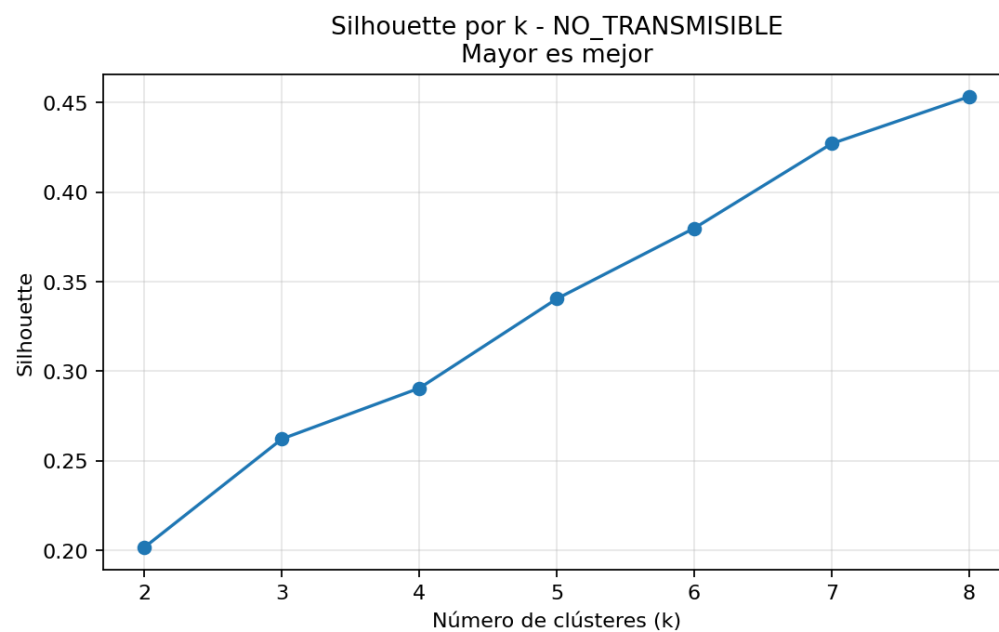
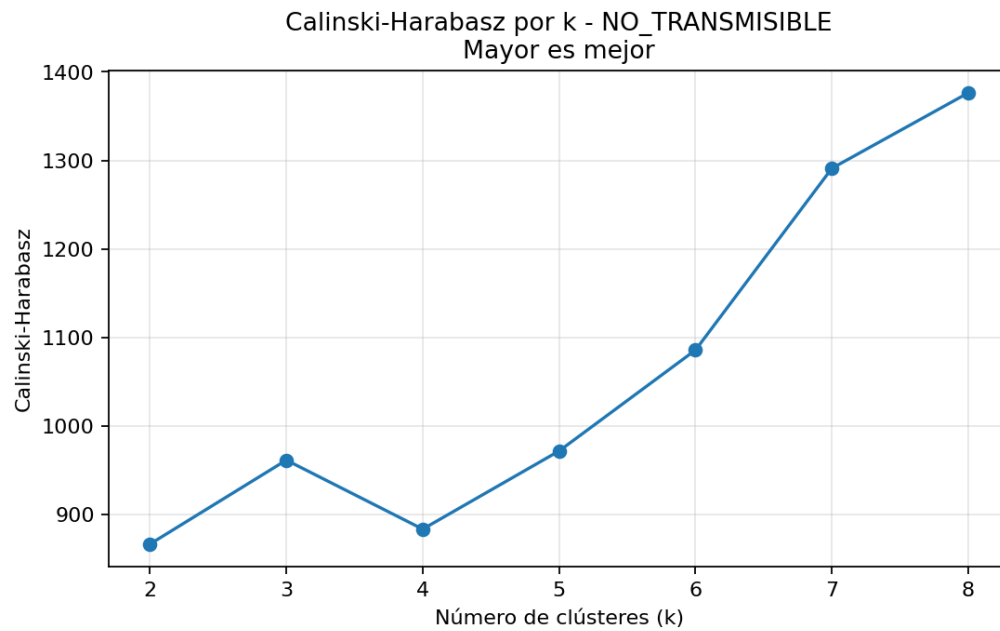


Figura 19

Índice Calinski–Harabasz según el número de grupos en el macrogrupo no transmisible.

**Figura 20**

Índice Davies–Bouldin según el número de grupos en el macrogrupo no transmisible.

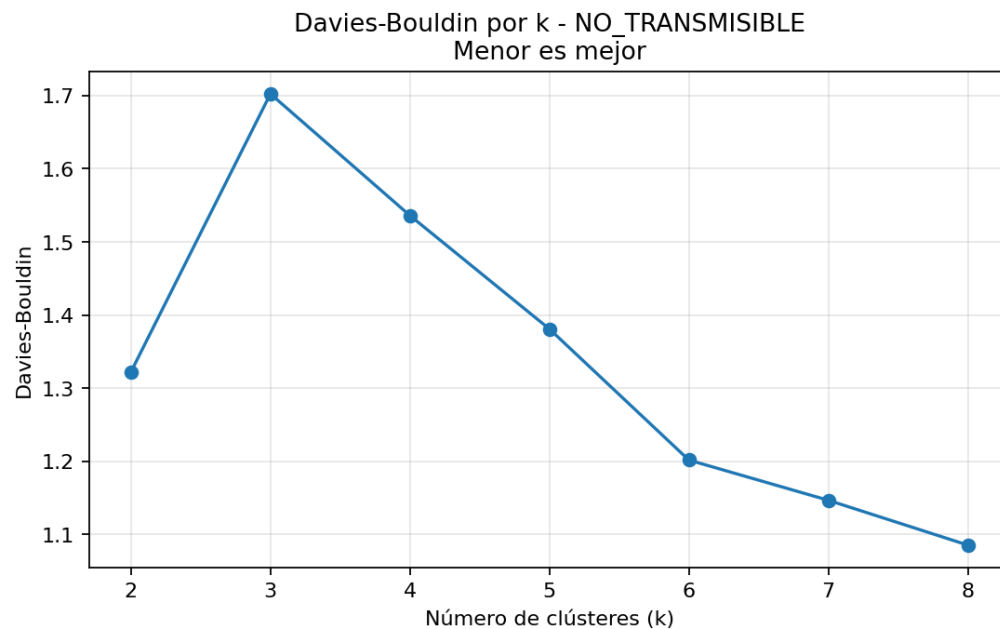
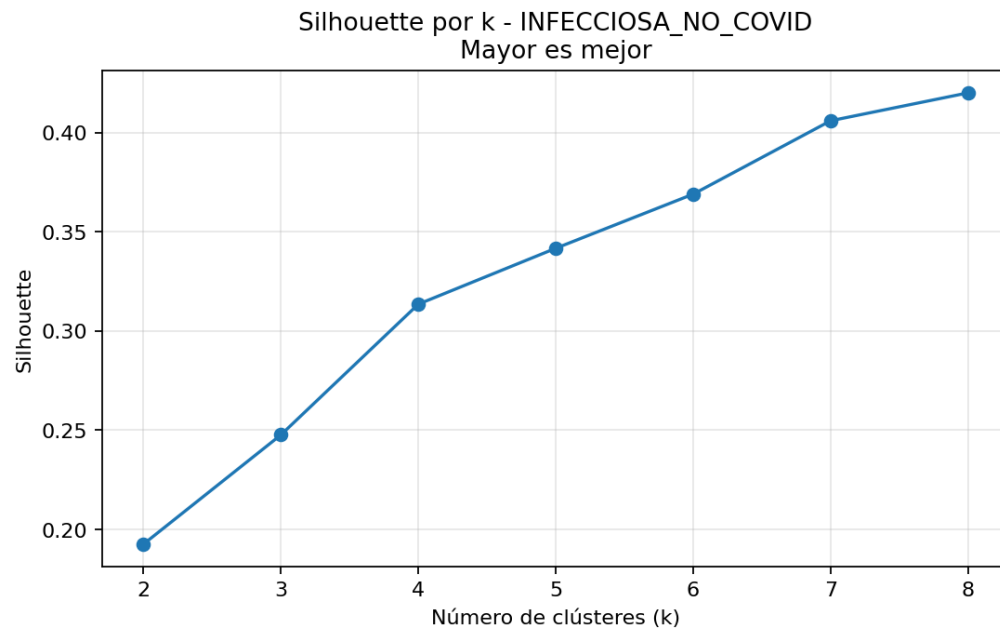


Figura 21

Coeficiente de silueta según el número de grupos en el macrogrupo infecciosa no COVID.

**Figura 22**

Índice Calinski-Harabasz según el número de grupos en el macrogrupo infecciosa no COVID.

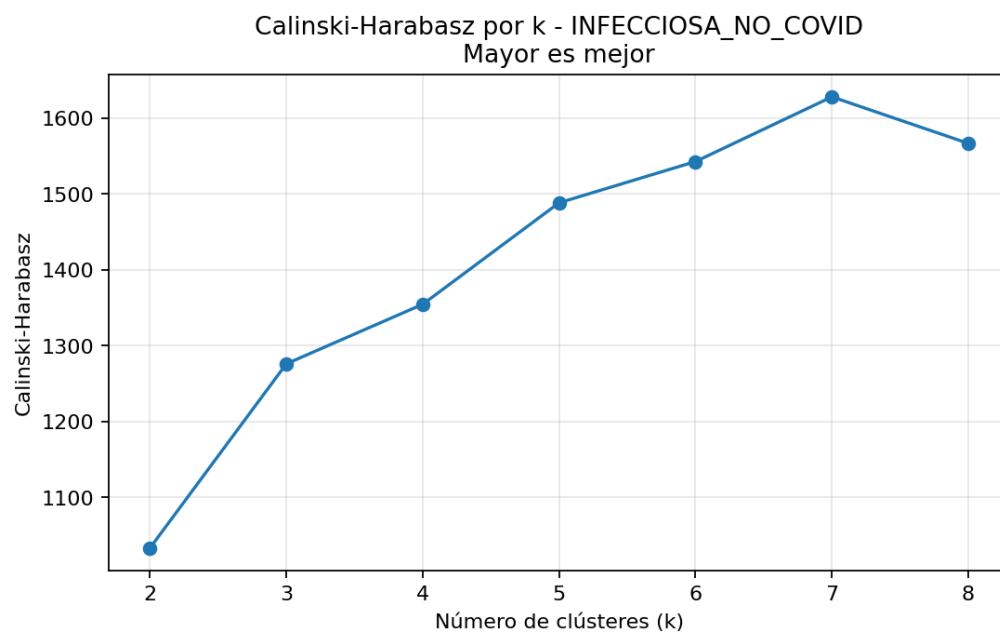
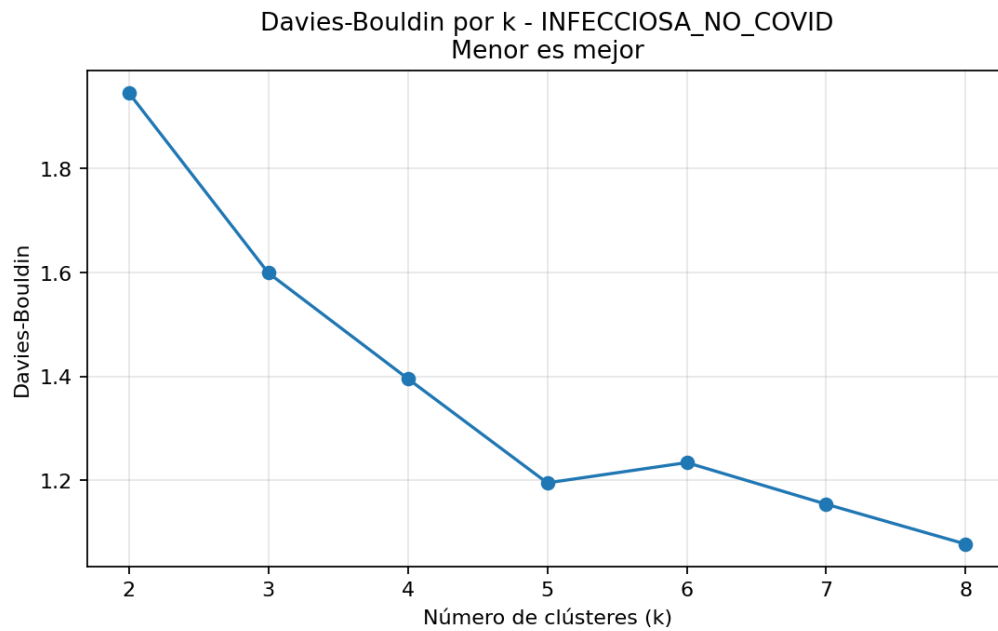


Figura 23

Índice Davies–Bouldin según el número de grupos en el macrogrupo infecciosa no COVID.



C.1.8 Estabilidad del agrupamiento

Figura 24

Estabilidad de la solución de agrupamiento mediante índice de Rand ajustado bajo cambios de semilla y remuestreo.

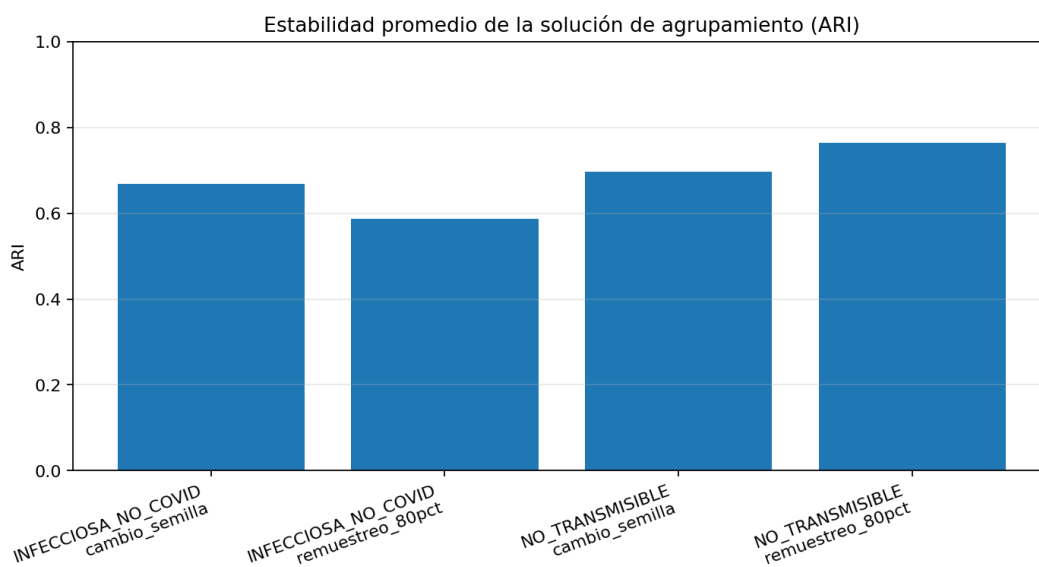
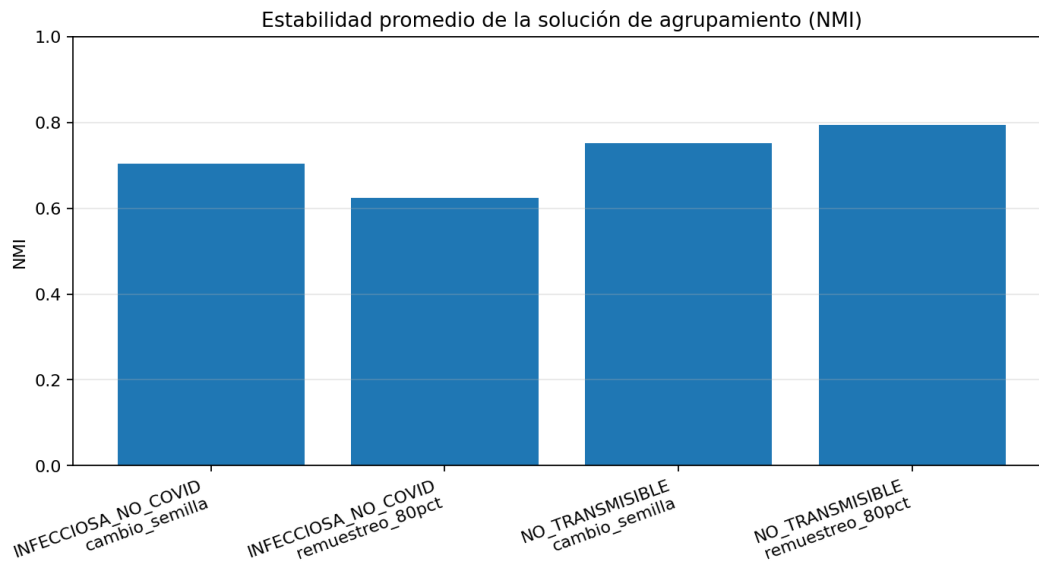


Figura 25

Estabilidad de la solución de agrupamiento mediante información mutua normalizada bajo cambios de semilla y remuestreo.



C.1.9 Evaluación preliminar de transformaciones financieras

Figura 26

Distribución del ingreso total original antes de aplicar transformaciones.

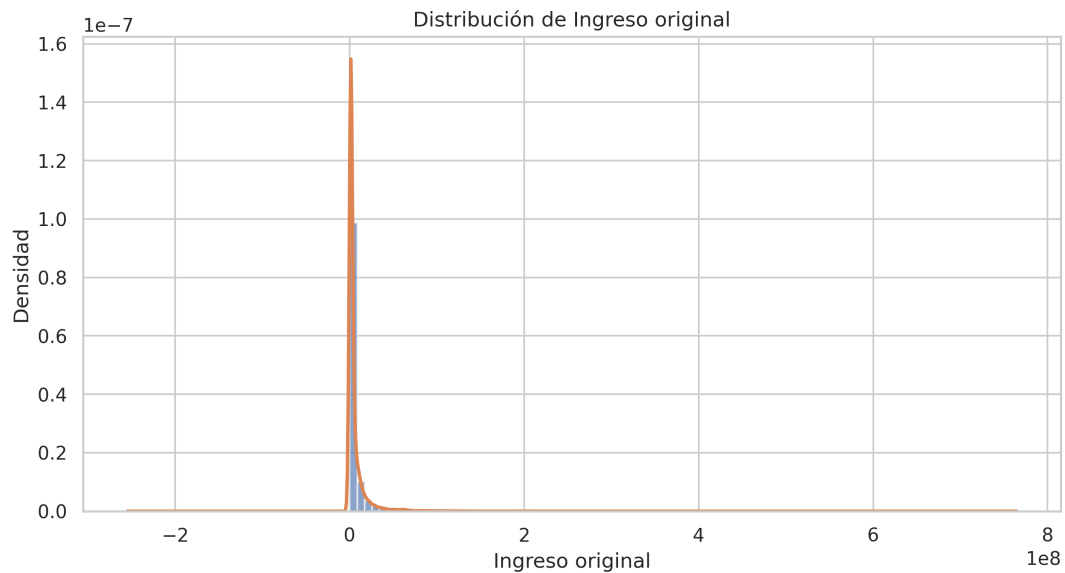


Figura 27

Comparación de dispersión y valores extremos entre transformaciones candidatas del ingreso total.

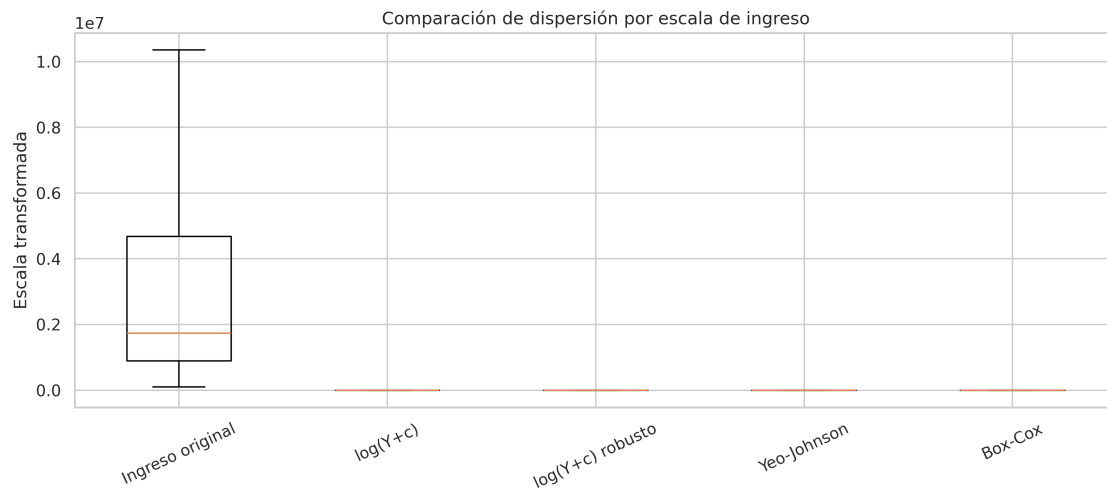
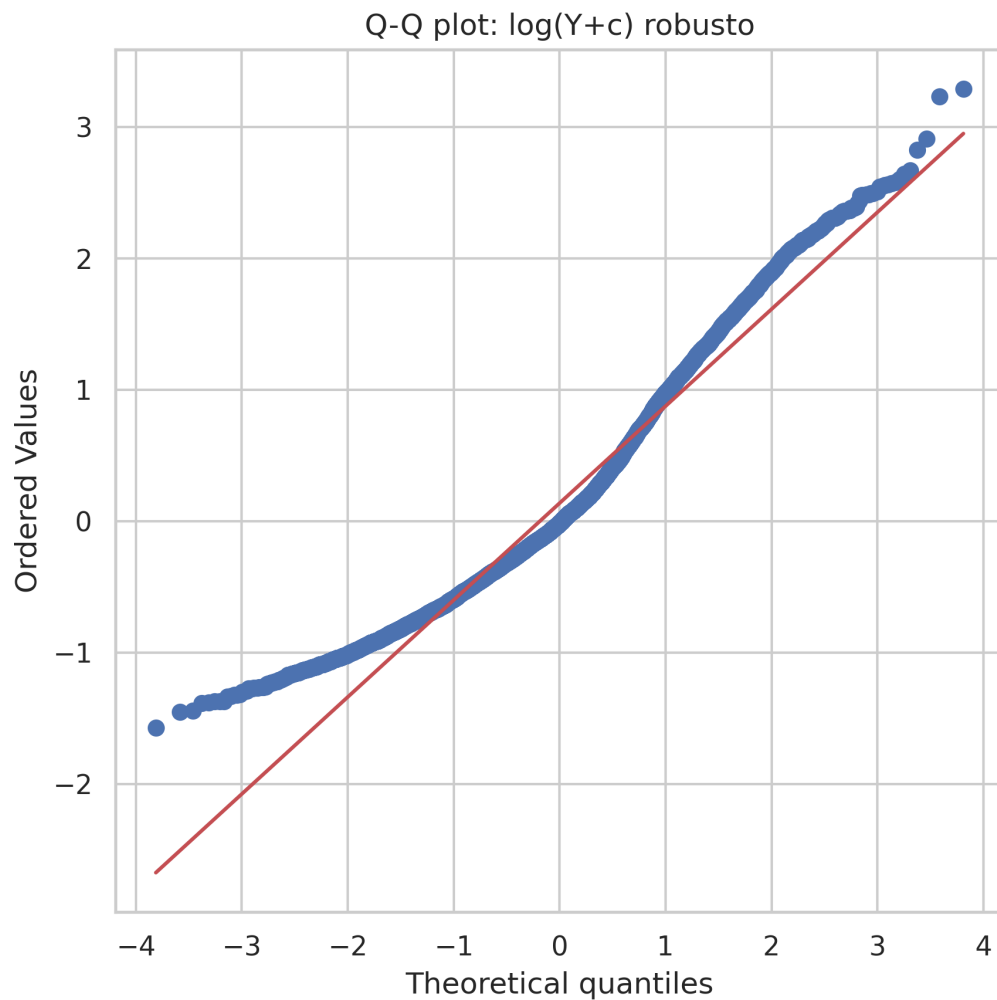


Figura 28

Gráfico Q-Q de la transformación logarítmica robusta del ingreso total.



Apéndice D. Complementos del Capítulo 6

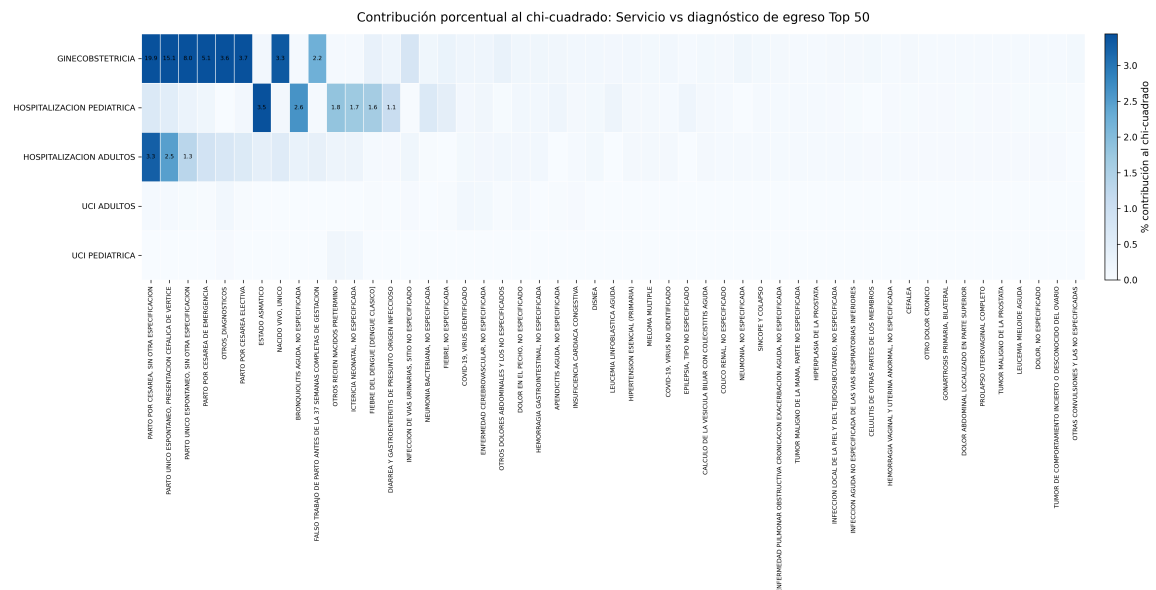
D.1 Figuras complementarias del análisis post hoc de asociaciones

Este complemento reúne las visualizaciones de contribución porcentual al estadístico chi-cuadrado correspondientes a los cruces post hoc priorizados en el Capítulo 6. Las figuras se trasladan fuera del cuerpo principal para mejorar la organización del capítulo, conservando sus títulos, etiquetas y función interpretativa.

D.1.1 Servicio y diagnóstico de egreso Top 50

Figura 29

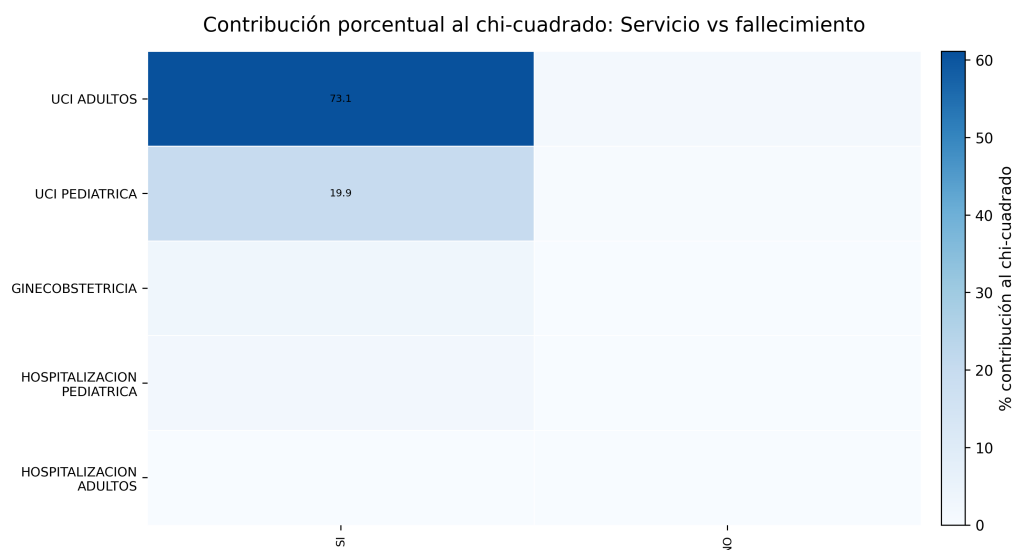
Contribución porcentual al chi-cuadrado en el cruce Servicio y diagnóstico de egreso Top 50.

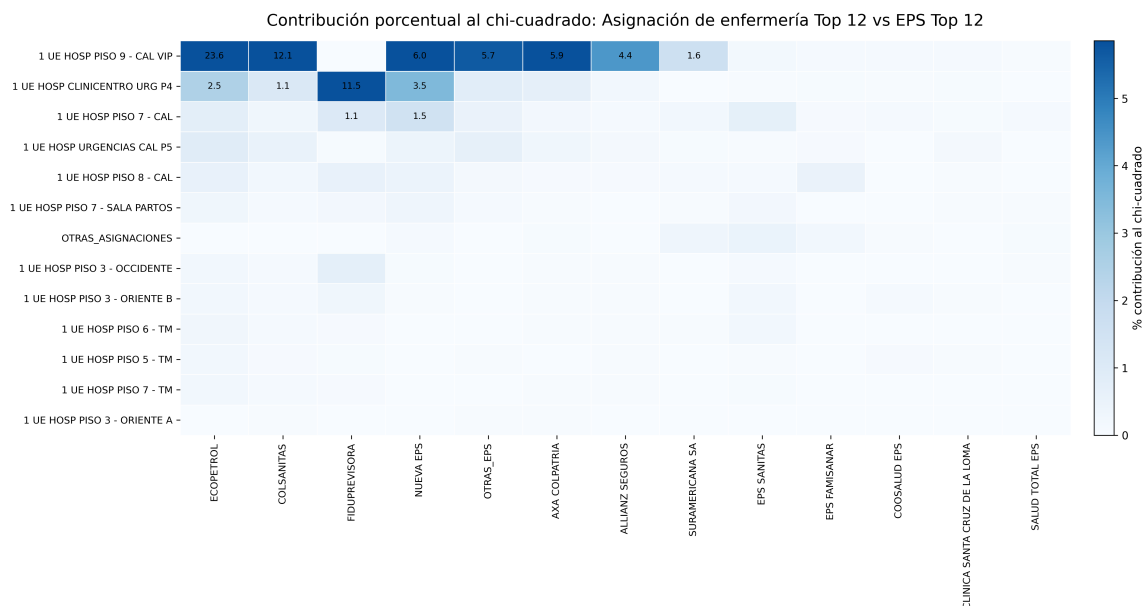


D.1.2 Servicio y fallecimiento

Figura 30

Contribución porcentual al chi-cuadrado en el cruce Servicio y fallecimiento.





Apéndice E. Complementos del Capítulo 7

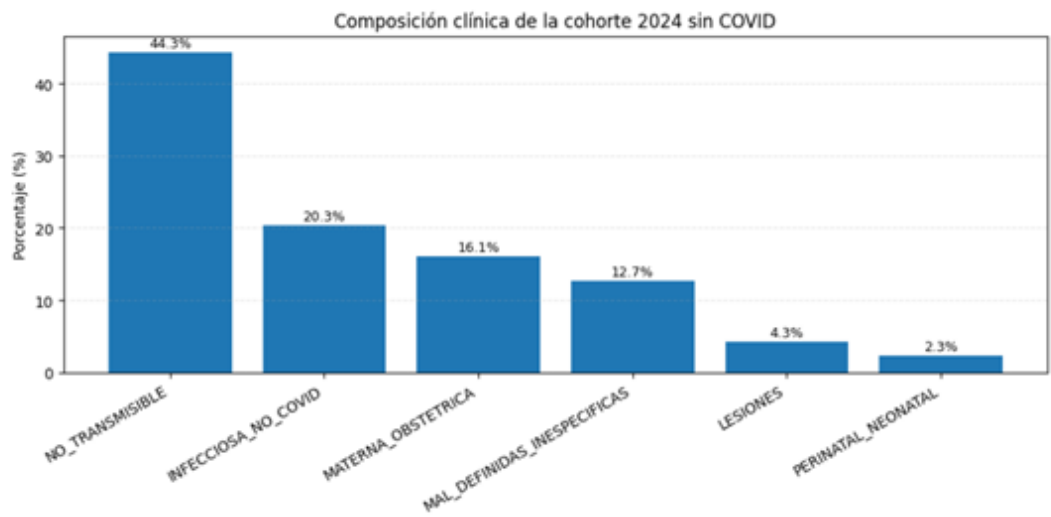
E.1 Propósito y criterio de organización del apéndice

Este apéndice reúne los elementos complementarios que respaldan el Capítulo 7 sin repetir la narrativa principal. El apéndice conserva evidencia necesaria para auditar la reclasificación, la selección de grupos, la mortalidad, la estabilidad, la calidad del dato y la interpretación clínica posterior. Para mantener claridad y lectura vertical, las tablas extensas se dividen en bloques temáticos y se excluyen listados de bajo valor interpretativo, como los Top 100 de diagnósticos raros y el detalle fila a fila de las 30 corridas de estabilidad.

E.2 Contexto clínico y auditoría de la reclasificación

E.2.1 Composición clínica de la cohorte 2024

Figura 33
Composición clínica de la cohorte 2024 según clasificación reconstruida



La Figura 33 contextualiza la distribución de macrogrupos clínicos antes de la construcción de los grupos finales. Su función es descriptiva; no constituye una métrica de calidad del agrupamiento ni una prueba de representatividad institucional.

Tabla 61

Principales diagnósticos en la categoría mal definidas e inespecíficas

Diagnóstico	Frecuencia
Fiebre, no especificada	617
Otros dolores abdominales y los no especificados	515
Dolor en el pecho, no especificado	244
Disnea	197
Otras convulsiones y las no especificadas	171
Síncope y colapso	168
Dolor abdominal localizado en parte superior	133
Cefalea	121
Otro dolor crónico	85
Náusea y vómito	84

Nota. La categoría indica inespecificidad diagnóstica del registro, no ausencia de importancia clínica.

E.2.2 Matriz de reclasificación original frente a reconstruida**Tabla 62**

Matriz de reclasificación: categorías reconstruidas COVID-19, infecciosas, lesiones y mal definidas

Macro original	COVID	Inf.	Les.	Mal def.	Total
Infecciosa	0	3.353	1	0	3.642
Lesiones	0	0	98	0	98
Mal definidas	14	60	918	3.274	4.785
Mat./per. previas	0	0	0	0	4.562
No transmisible	0	1.917	116	38	13.168
Otra/desc.	0	0	0	16	25
Total reconstr.	14	5.330	1.133	3.328	26.280

Nota. El total original se conserva para lectura por fila. Las columnas materna, no transmisible y perinatal se presentan en la Tabla 63 para evitar formato horizontal.

Tabla 63
Matriz de reclasificación: categorías reconstruidas materna, no transmisible y perinatal

Macro original	Materna	No trans.	Perinatal	Total
Infecciosa	0	288	0	3.642
LESIONES	0	0	0	98
Mal definidas	263	256	0	4.785
Mat./per. previas	3.957	0	605	4.562
No transmisible	0	11.094	3	13.168
Otra/desc.	9	0	0	25
Total reconstr.	4.229	11.638	608	26.280

Nota. Las dos tablas de matriz permiten auditar cambios de clasificación sin usar páginas horizontales. La suma de categorías reconstruidas coincide con 26.280 episodios de 2024 antes de excluir COVID-19.

E.3 Complementos metodológicos de agrupamiento

E.3.1 Sensibilidad al número de componentes SVD

Tabla 64
Sensibilidad SVD en no transmisible

Comp.	Var.	No retenida	Silueta	DB	ARI vs final	Decisión
5	0,235	0,765	0,483	0,867	0,789	Sensibilidad
10	0,398	0,602	0,260	1,716	1,000	Configuración final
20	0,591	0,409	0,142	1,886	0,337	Sensibilidad
30	0,695	0,305	0,073	2,691	0,135	Sensibilidad

Nota. La configuración de 10 componentes se conserva como configuración principal parsimoniosa; no representa un número óptimo universal.

Tabla 65*Sensibilidad SVD en infecciosa no COVID*

Comp.	Var.	No retenida	Silueta	DB	ARI vs final	Decisión
5	0,524	0,476	0,382	1,272	0,905	Sensibilidad
10	0,689	0,311	0,248	1,598	1,000	Configuración final
20	0,856	0,144	0,250	1,893	0,283	Sensibilidad
30	0,953	0,047	0,222	2,033	0,272	Sensibilidad

Nota. El análisis documenta sensibilidad de la solución ante cambios de dimensionalidad; no equivale a búsqueda de hiperparámetro óptimo formal.

E.3.2 Macrogrupos no subdivididos

Tabla 66*Evaluación resumida de macrogrupos conservados como grupos únicos*

Macrogrupo	<i>n</i>	<i>k</i> eval.	<i>k</i> mayor sil.	Silueta máx.	Tamaño mín.
LESIONES	1.133	2–6	5	0,361	24
mal definidas o inespecíficas	3.328	2–6	2	0,335	1.500
materna-obstétrica	4.229	2–6	6	0,302	257
perinatal-neonatal	608	2–6	6	0,797	12

Nota. La subdivisión de estos macrogrupos se trató como exploratoria. La solución final los conservó como grupos únicos; por eso no aplica reportar ARI/NMI de una partición interna final.

Tabla 67*Criterio de conservación como grupo único*

Macrogrupo	Decisión metodológica
LESIONES	Conservar como grupo único en el capítulo; la subdivisión mostró fragmentación de tamaño mínimo.
mal definidas o inespecíficas	Conservar como grupo único; la finalidad principal fue auditoría de calidad del dato, no segmentación clínica interna.
materna-obstétrica	Conservar como grupo único; la lectura institucional favorece perfil obstétrico agregado y evita fragmentación.
perinatal-neonatal	Conservar como grupo único; aunque la silueta fue alta, el tamaño mínimo exploratorio fue muy bajo.

Nota. La tabla formula la decisión sin presentar los resultados exploratorios como solución final.

E.3.3 Estabilidad de la solución

Tabla 68*Resumen de estabilidad mediante ARI con 30 corridas*

Macrogrupo	Validación	Corridas	Media	DE	Mín.
infecciosa no COVID	Cambio semilla	30	0,644	0,194	0,218
infecciosa no COVID	Remuestreo	30	0,562	0,183	0,266
	80 %				
no transmisible	Cambio semilla	30	0,705	0,145	0,352
no transmisible	Remuestreo	30	0,675	0,231	0,265
	80 %				

Nota. El resumen conserva los estadísticos necesarios para interpretar reproducibilidad. El detalle fila a fila se excluye por extensión y redundancia.

Tabla 69*Resumen de estabilidad mediante NMI con 30 corridas*

Macrogrupo	Validación	Corridas	Media	Mín.	Máx.
infecciosa no COVID	Cambio semilla	30	0,669	0,299	0,927
infecciosa no COVID	Remuestreo	30	0,605	0,340	0,899
	80 %				
no transmisible	Cambio semilla	30	0,744	0,569	0,922
no transmisible	Remuestreo	30	0,741	0,396	1,000
	80 %				

Nota. NMI complementa ARI al cuantificar similitud de información entre particiones.

E.4 Mortalidad, ajuste por edad y discrepancia clínica

E.4.1 Mortalidad cruda por macrogrupo

Tabla 70*Mortalidad cruda por macrogrupo con IC95 %*

Macrogrupo	<i>n</i>	Fallec.	Mort. %	IC95 inf.	IC95 sup.
infecciosa no COVID	5.330	121	2,27	1,90	2,71
no transmisible	11.638	264	2,27	2,01	2,56
mal definidas o inespecíficas	3.328	59	1,77	1,38	2,28
LESIONES	1.133	19	1,68	1,08	2,60
perinatal-neonatal	608	6	0,99	0,45	2,14
materna-obstétrica	4.229	0	0,00	0,00	0,09

Nota. IC95 % binomial. La mortalidad por macrogrupo complementa la lectura por grupo clínico final del capítulo principal.

Tabla 71*Comparación de mortalidad por macrogrupo frente al resto de la cohorte*

Macrogrupo	Aporte muertes %	OR crudo	<i>p</i> Holm
infecciosa no COVID	25,80	1,37	0,0111
no transmisible	56,29	1,63	< 0,001
mal definidas o inespecíficas	12,58	0,99	1,0000
LESIONES	4,05	0,94	1,0000
perinatal-neonatal	1,28	0,54	0,3972
materna-obstétrica	0,00	0,00	< 0,001

Nota. Las comparaciones son grupo frente al resto y los valores *p* están ajustados con Holm.

E.4.2 Mortalidad ajustada por edad y sexo**Tabla 72***Mortalidad ajustada por edad y sexo por grupo clínico final*

Grupo clínico final	<i>n</i>	Fallec.	Cruda %	Ajustada %	IC95 inf.	IC95 sup.
Perinatal-neonatal C01	608	6	0,99	4,35	1,57	8,90
Infecc. no COVID C01	1.704	56	3,29	4,07	3,07	5,13
Infecc. no COVID C02	1.803	43	2,38	2,67	1,96	3,35
No transmisible C02	2.660	107	4,02	2,14	1,74	2,59
No transmisible C03	7.114	139	1,95	2,08	1,75	2,43
Mal definidas C01	3.328	59	1,77	1,42	1,08	1,80
Lesiones C01	1.133	19	1,68	1,40	0,84	2,04
Infecc. no COVID C03	1.823	22	1,21	1,08	0,67	1,45
No transmisible C01	1.864	18	0,97	0,87	0,58	1,28
Materna-obstétrica C01	4.229	0	0,00	0,11	0,10	0,17

Nota. Método: regresión logística L2 con estandarización marginal y 200 remuestreos bootstrap para IC95 %.

Tabla 73*Modelo logístico ajustado por edad y sexo*

Término	OR	IC95 inf.	IC95 sup.	p
Intercepto	0,007	0,004	0,010	< 0,001
Infecciosa no COVID C02	0,609	0,385	0,966	0,0349
Infecciosa no COVID C03	0,224	0,133	0,375	< 0,001
Lesiones C01	0,310	0,180	0,535	< 0,001
Mal definidas o inespecíficas C01	0,327	0,220	0,485	< 0,001
Materna-obstétrica C01	0,000	0,000	–	0,9987
No transmisible C01	0,192	0,111	0,333	< 0,001
No transmisible C02	0,492	0,342	0,707	< 0,001
No transmisible C03	0,481	0,342	0,678	< 0,001
Perinatal-neonatal C01	1,448	0,584	3,587	0,4245
Sexo: mujer	1,050	0,851	1,295	0,6490
Edad numérica	1,034	1,029	1,039	< 0,001

Nota. La categoría de referencia corresponde al grupo omitido por codificación del modelo. El intercepto se reporta por completitud del modelo y no debe interpretarse como comparación directa entre grupos.

La mortalidad ajustada y el modelo logístico responden a la observación sobre confusión por edad y sexo. El análisis no ajusta por comorbilidad porque la base disponible no contiene diagnósticos secundarios ni índices Charlson o Elixhauser.

E.4.3 Discrepancia clínica**Tabla 74***Mortalidad según presencia de discrepancia clínica*

Grupo	<i>n</i>	Fallec.	Mortalidad %	IC95 inf.	IC95 sup.
Sin discrepancia clínica	22.381	407	1,82	1,65	2,00
Con discrepancia clínica	3.885	62	1,60	1,25	2,04

Nota. Prueba chi-cuadrado 2x2; $p = 0,3334$. No se observó evidencia estadística de diferencia de mortalidad por discrepancia clínica.

Tabla 75*Discrepancia clínica por macrogrupo*

Macrogrupo	<i>n</i>	Discrepancias	Discrepancia %
Lesiones	1.133	1.035	91,35
Infecciosa no COVID	5.330	1.977	37,09
Materna-obstétrica	4.229	272	6,43
No transmisible	11.638	544	4,67
Mal definidas o inespecíficas	3.328	54	1,62
Perinatal-neonatal	608	3	0,49

Nota. La discrepancia no se distribuye uniformemente; se concentra especialmente en lesiones e infecciosas no COVID.

Tabla 76*Discrepancia clínica por grupo clínico final*

Grupo clínico final	<i>n</i>	Discrepancias	Discrepancia %
Infecciosa no COVID C03	1.823	1.781	97,70
Lesiones C01	1.133	1.035	91,35
Infecciosa no COVID C01	1.704	134	7,86
No transmisible C03	7.114	521	7,32
Materna-obstétrica C01	4.229	272	6,43
Infecciosa no COVID C02	1.803	62	3,44
Mal definidas o inespecíficas C01	3.328	54	1,62
No transmisible C02	2.660	23	0,86
Perinatal-neonatal C01	608	3	0,49
No transmisible C01	1.864	0	0,00

Nota. La tabla responde a la observación sobre concentración de discrepancias entre grupos clínicos finales.

Estos resultados muestran que la discrepancia clínica no se comporta como una variable homogéneamente distribuida. Por ello, el análisis la conserva como control de calidad y trazabilidad de reclasificación, no como variable de agrupamiento.

E.5 Caracterización sintética y paneles descriptivos

E.5.1 Etiquetas clínicas interpretativas de los grupos finales

La Tabla 77 documenta los nombres interpretativos asignados a cada grupo clínico final. Estas etiquetas se construyeron con información observada después del agrupamiento: macrogrupo, diagnóstico dominante, edad, sexo, mortalidad, aporte a muertes y concentración de discrepancia clínica. No constituyen diagnósticos nuevos ni modifican los identificadores computacionales de la base; su función es mejorar la lectura clínica y evitar que identificadores aislados como C01, C02 o C03 aparezcan como nombres vacíos.

Tabla 77*Etiquetas interpretativas de los subgrupos clínicos finales*

Grupo clínico final	Etiqueta interpretativa
No transmisible C01	No transmisible digestivo-hemorrágico de baja mortalidad.
No transmisible C02	No transmisible de adulto mayor cardiovascular-neurológico de alta mortalidad.
No transmisible C03	No transmisible heterogéneo de alto volumen y aporte al total de fallecimientos.
Infecciosa no COVID C01	Infecciosa no COVID masculina, predominantemente dengue/gastroenteritis/neumonía, con mayor mortalidad ajustada por edad y sexo.
Infecciosa no COVID C02	Infecciosa no COVID femenina, predominantemente dengue/gastroenteritis/neumonía, con mortalidad intermedia ajustada por edad y sexo.
Infecciosa no COVID C03	Infecciosa no COVID urinaria y de tejidos blandos, con menor mortalidad cruda y alta reclasificación.
Mal definidas o inespecíficas C01	Sintomático inespecífico o mal definido, prioritario para auditoría de codificación.
Lesiones C01	Lesiones y trauma hospitalario.
Materna-obstétrica C01	Materna-obstétrica sin muertes observadas en la cohorte.
Perinatal-neonatal C01	Perinatal-neonatal de bajo volumen y estimación de mortalidad imprecisa.

Nota. La etiqueta resume patrones dominantes y no reemplaza la revisión clínica individual. La base conserva los identificadores originales completos para trazabilidad.

Tabla 78*Criterio principal utilizado para nombrar cada grupo*

Grupo clínico final	Criterio de denominación
No transmisible C01	Baja mortalidad cruda y perfil no transmisible compatible con diagnóstico digestivo-hemorrágico dominante en el análisis de sensibilidad sin CIE-3 raros.
No transmisible C02	Edad mediana alta, diagnósticos cardiovasculares y neurológicos dominantes, y mayor mortalidad cruda local.
No transmisible C03	Mayor tamaño del capítulo y mayor aporte absoluto a muertes, con diagnósticos no transmisibles heterogéneos.
Infecciosa no COVID C01	Separación masculina, predominio de dengue/gastroenteritis/neumonía y mortalidad ajustada más alta.
Infecciosa no COVID C02	Separación femenina, predominio de dengue/gastroenteritis/neumonía y mortalidad intermedia ajustada por edad y sexo.
Infecciosa no COVID C03	Menor mortalidad cruda dentro de infecciosas y concentración muy alta de discrepancia clínica, con perfil urinario y tejidos blandos.
Mal definidas o inespecíficas C01	Diagnósticos consignados como signos, síntomas o condiciones inespecíficas, útiles para auditoría de codificación.
Lesiones C01	Macrogrupo único de lesiones, trauma y causas externas.
Materna-obstétrica C01	Macrogrupo único materno-obstétrico, sin muertes observadas en 2024.
Perinatal-neonatal C01	Macrogrupo único perinatal-neonatal, con bajo volumen y amplia incertidumbre en mortalidad.

Nota. Cuando el criterio usa resultados de sensibilidad, la etiqueta se interpreta como apoyo descriptivo, no como cambio de la solución principal.

*E.5.2 Perfil sintético de grupos prioritarios***Tabla 79***Perfil cuantitativo de grupos priorizados*

Grupo clínico final	<i>n</i>	Mort. %	IC95 inf.	IC95 sup.	Mediana edad	P25–P75
No transmisible C02	2.660	4,0	3,3	4,8	79	65–85
Infecciosa no COVID C01	1.704	3,3	2,5	4,2	21	7–65
Infecciosa no COVID C02	1.803	2,4	1,8	3,2	27	10–69
No transmisible C03	7.114	2,0	1,7	2,3	53	29–67
Mal definidas o inespecíficas C01	3.328	1,8	1,4	2,3	62	35–76

Nota. La priorización se basa en mortalidad cruda, incertidumbre, tamaño del grupo, aporte a muertes o relevancia de auditoría diagnóstica.

Tabla 80*Atributos dominantes de grupos priorizados*

Grupo clínico final	Sexo top	CIE-3 top	Diagnóstico top
No transmisible C02	Mujer	I67	Enfermedad cerebrovascular, no especificada (14,0 %).
Infecciosa no COVID C01	Hombre	A90	Fiebre del dengue [dengue clásico] (34,5 %).
Infecciosa no COVID C02	Mujer	A90	Fiebre del dengue [dengue clásico] (39,8 %).
No transmisible C03	Mujer	M54	Cólico renal, no especificado (2,2 %).
Mal definidas o inespecíficas C01	Mujer	R10	Fiebre, no especificada (18,5 %).

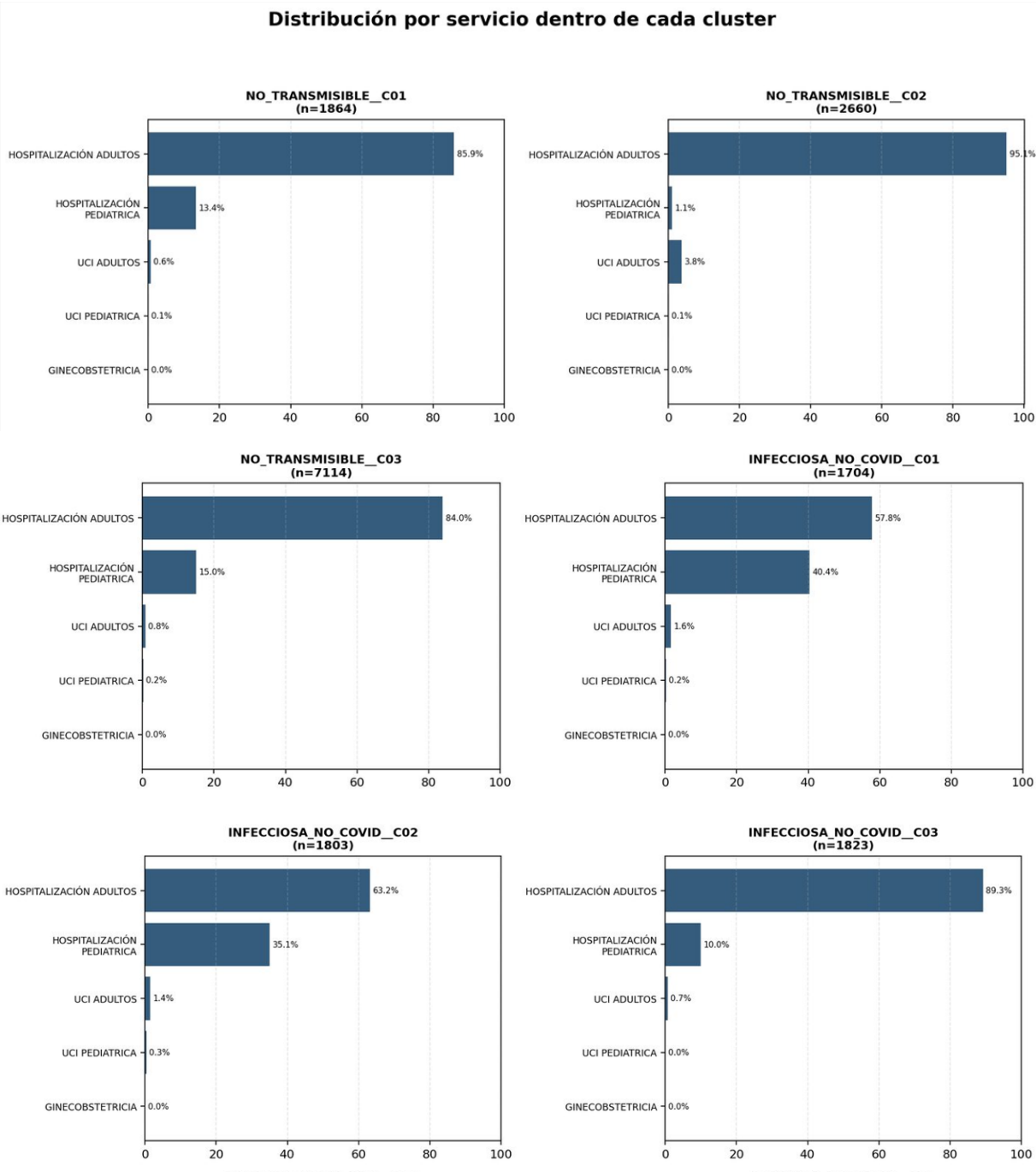
Nota. La tabla reemplaza listados extensos de diagnósticos, servicios y unidades médicas. La información visual completa se conserva mediante los paneles .

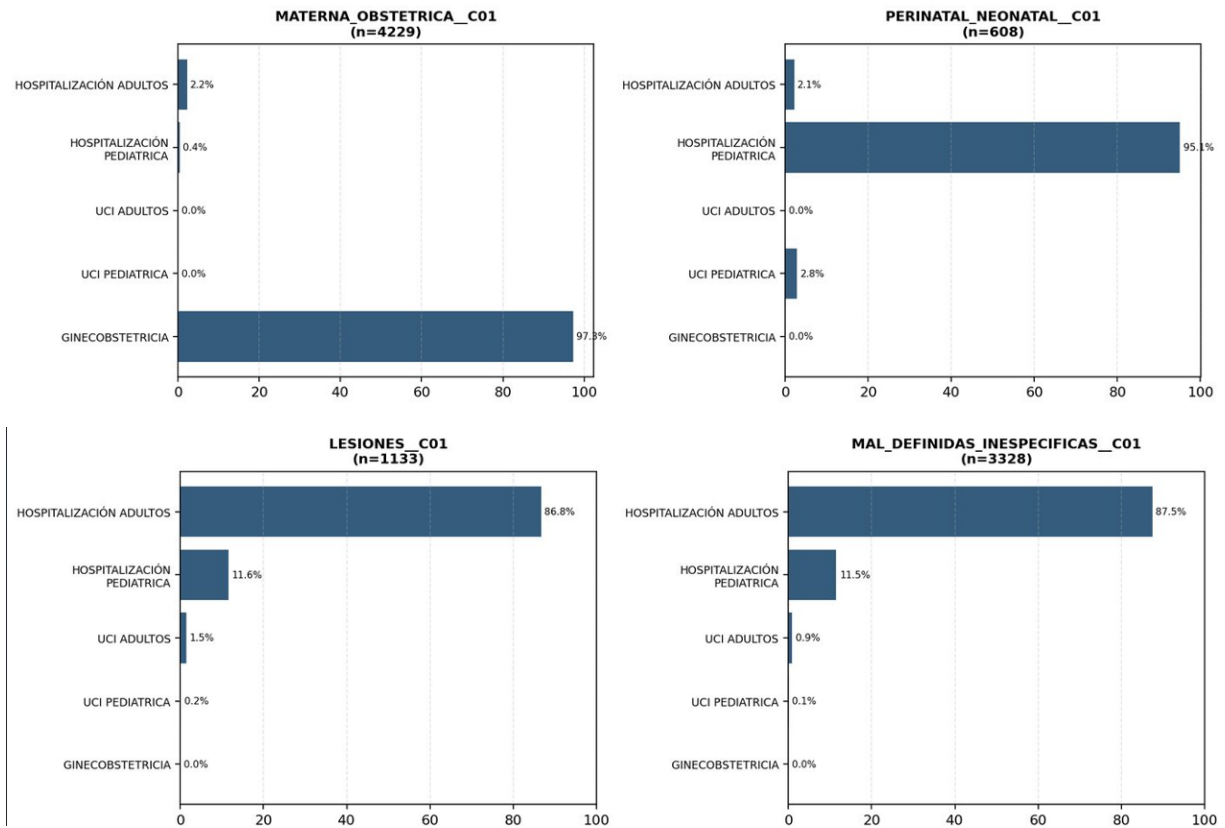
E.5.3 Paneles descriptivos por grupo clínico final

Los siguientes paneles se conservan porque la revisión señaló que las figuras mencionadas mediante los paneles debían estar presentes; sin ellas, la interpretación de los grupos quedaba sin soporte visual. Estos paneles cumplen una función descriptiva y de plausibilidad clínica posterior; no constituyen métricas de validación algorítmica.

E.5.3.1 Distribución por servicio hospitalario. La distribución por servicio permite evaluar si los grupos se concentran en áreas asistenciales compatibles con su perfil clínico. Esta variable no formó parte de la construcción de los grupos.

Figura 34*Distribución por servicio dentro de cada grupo clínico final*





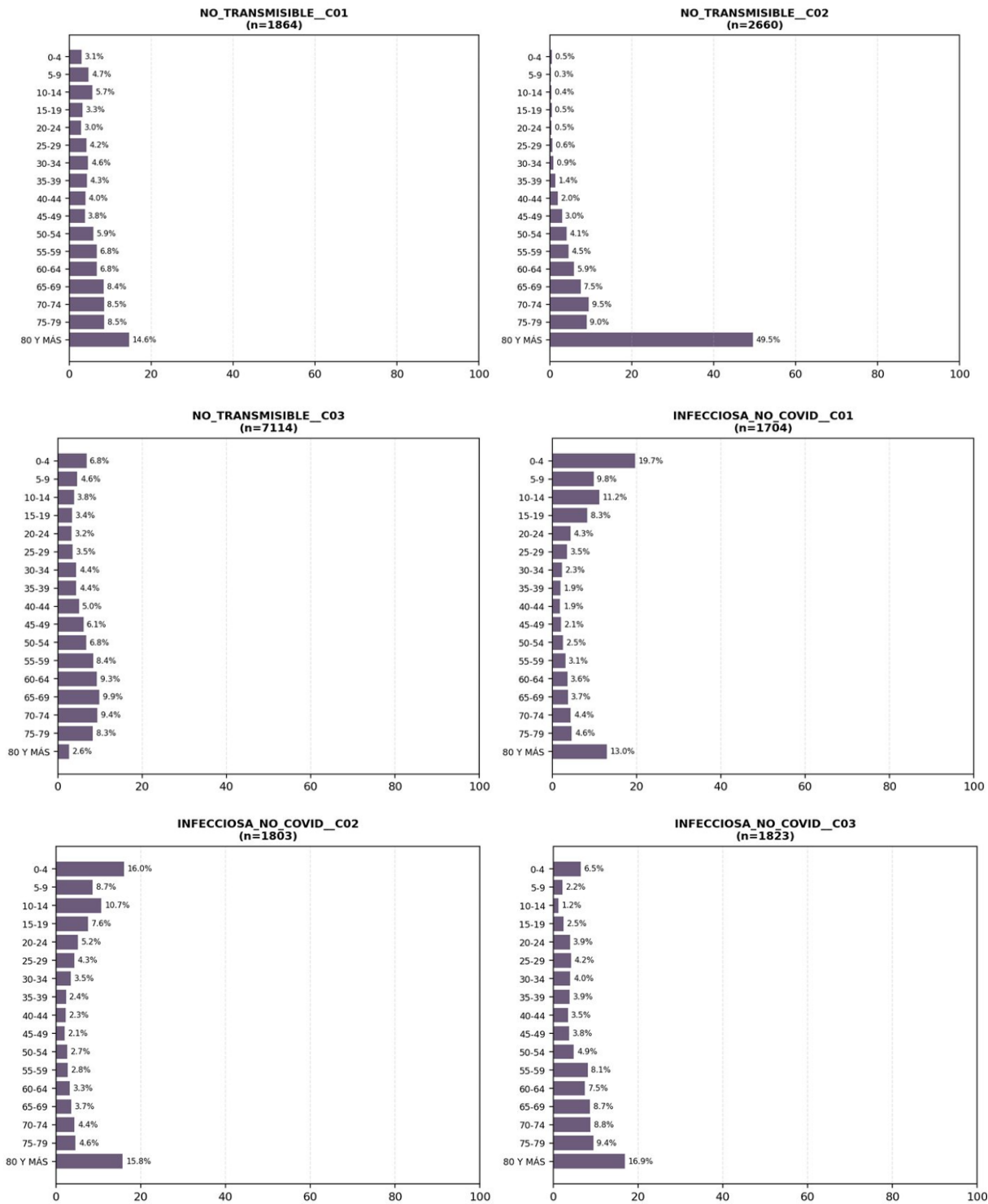
La distribución por servicio permite evaluar si los grupos se concentran en áreas asistenciales compatibles con su perfil clínico. Esta variable no formó parte de la construcción de los grupos.

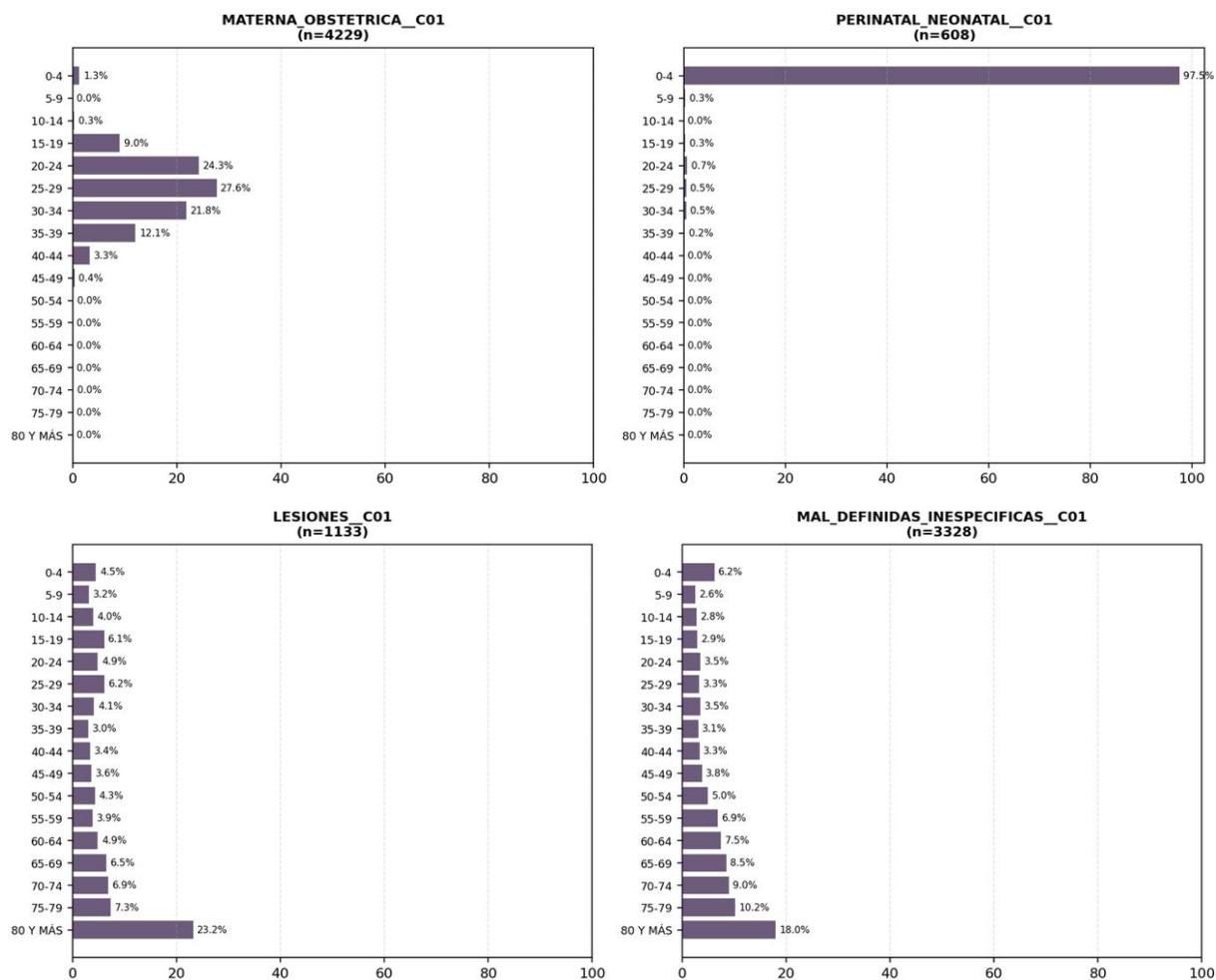
E.5.3.2 Distribución por grupo de edad. La edad permite diferenciar perfiles pediátricos, obstétricos, adultos y de adulto mayor. El capítulo principal reporta medidas de tendencia central y dispersión para evitar depender solo de la visualización.

Figura 35

Distribución por grupo de edad dentro de cada grupo clínico final

Distribución por grupo de edad dentro de cada cluster



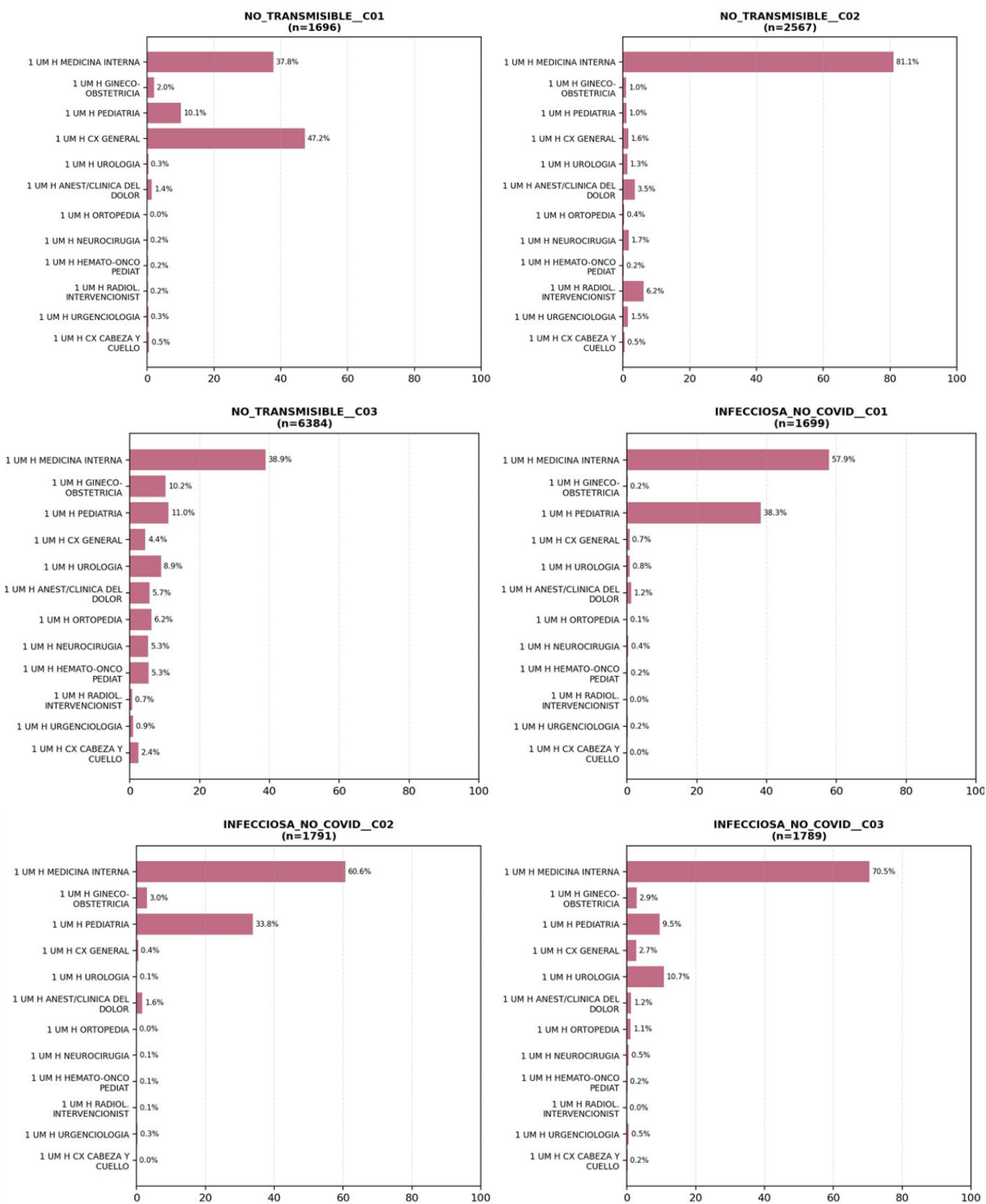


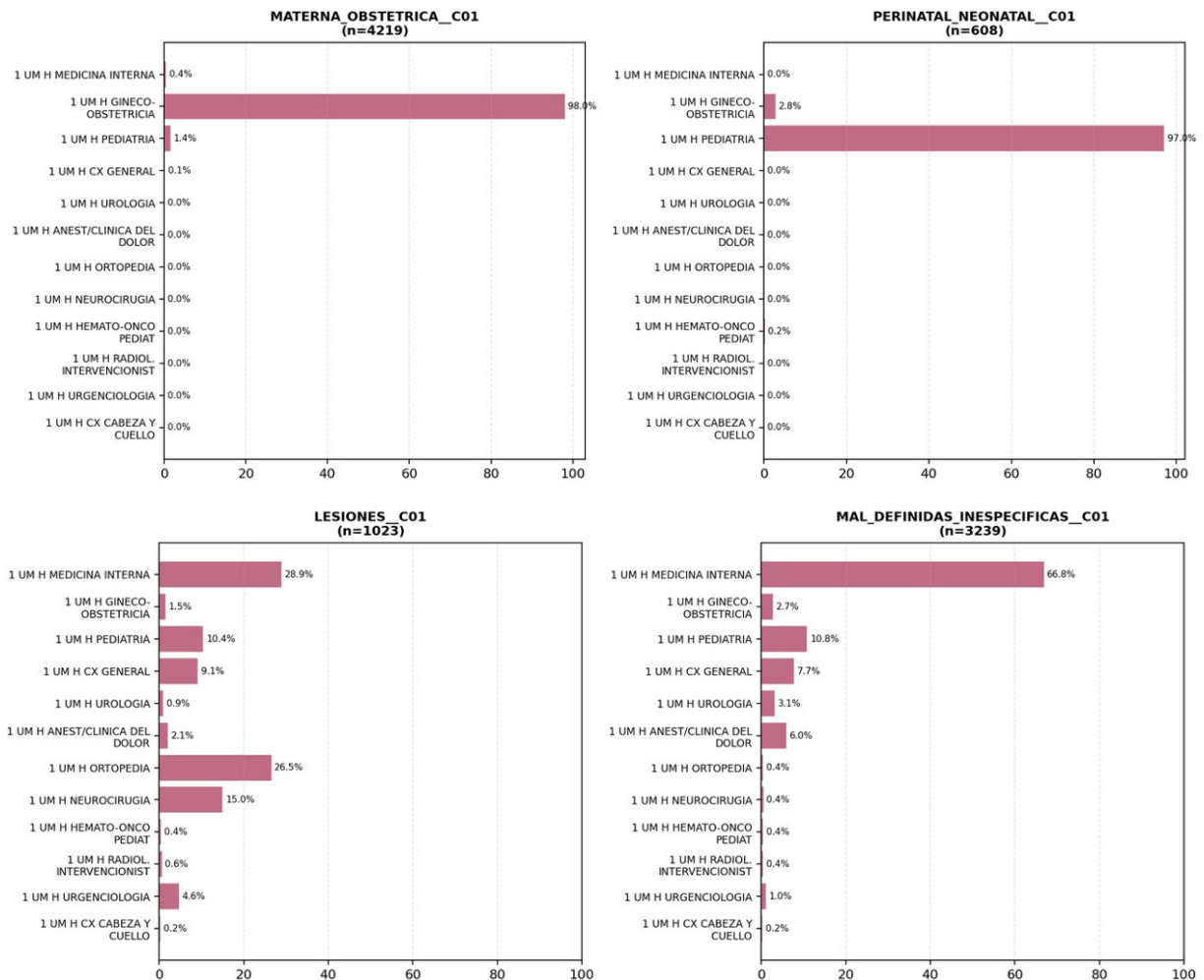
E.5.3.3 Distribución por unidad médica. La unidad médica se analiza después del agrupamiento. Su inclusión permite revisar la relación entre grupos clínicos y organización asistencial sin introducir circularidad en la formación de los grupos.

Figura 36

Distribución por unidad médica dentro de cada grupo clínico final

Distribución por UO médica dentro de cada cluster



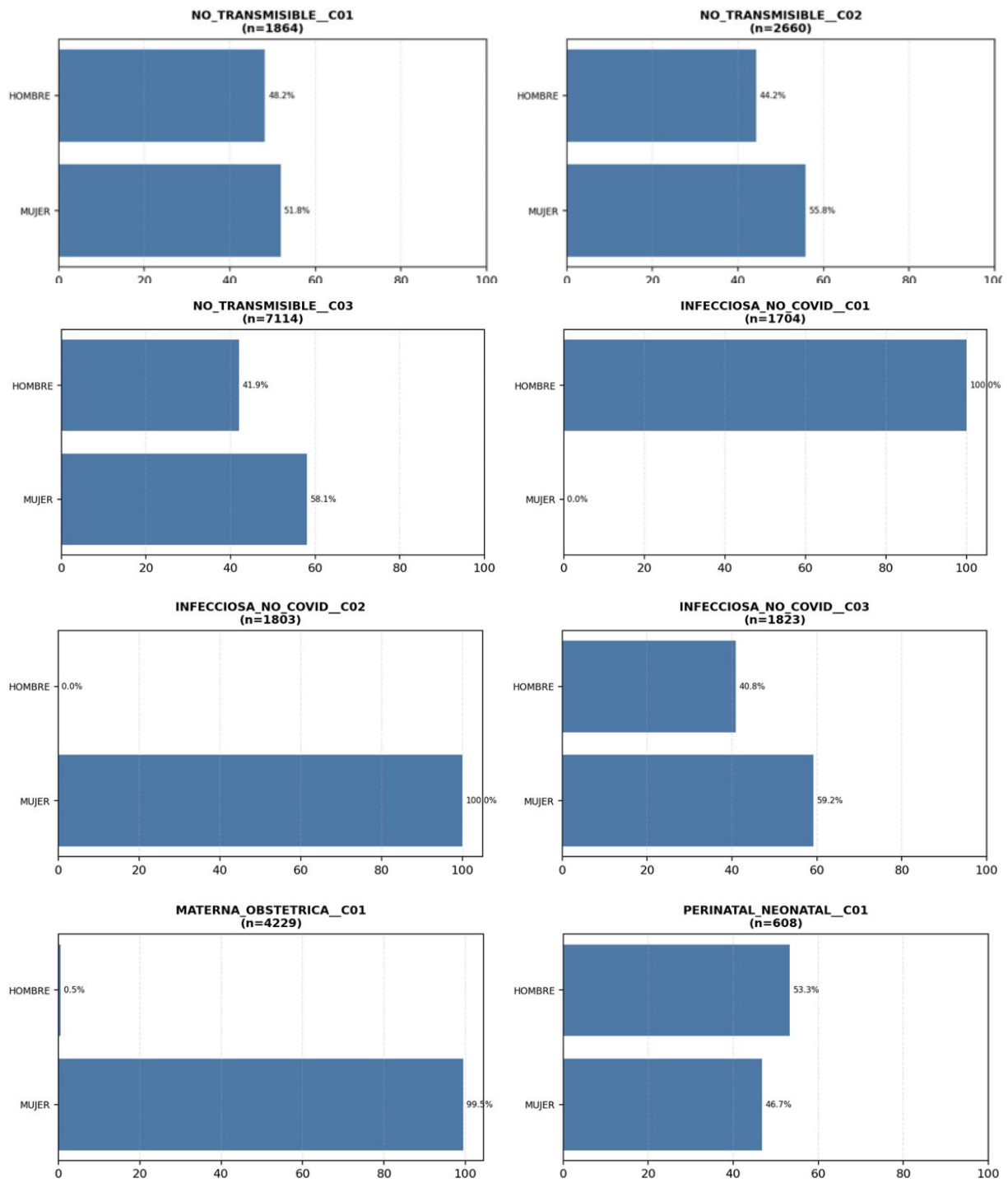


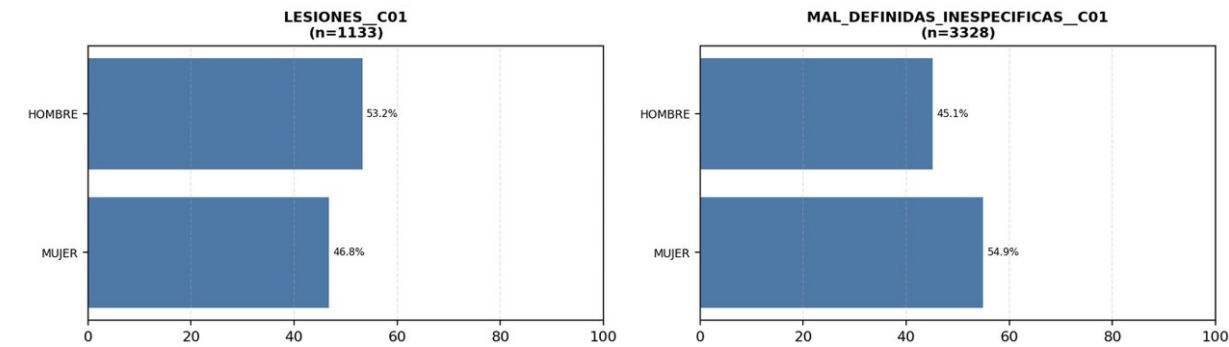
E.5.3.4 Distribución por sexo. La distribución por sexo permite describir composición demográfica. En no transmisible, el algoritmo final no incluyó sexo para privilegiar estructura diagnóstica y edad.

Figura 37

Distribución por sexo dentro de cada grupo clínico final

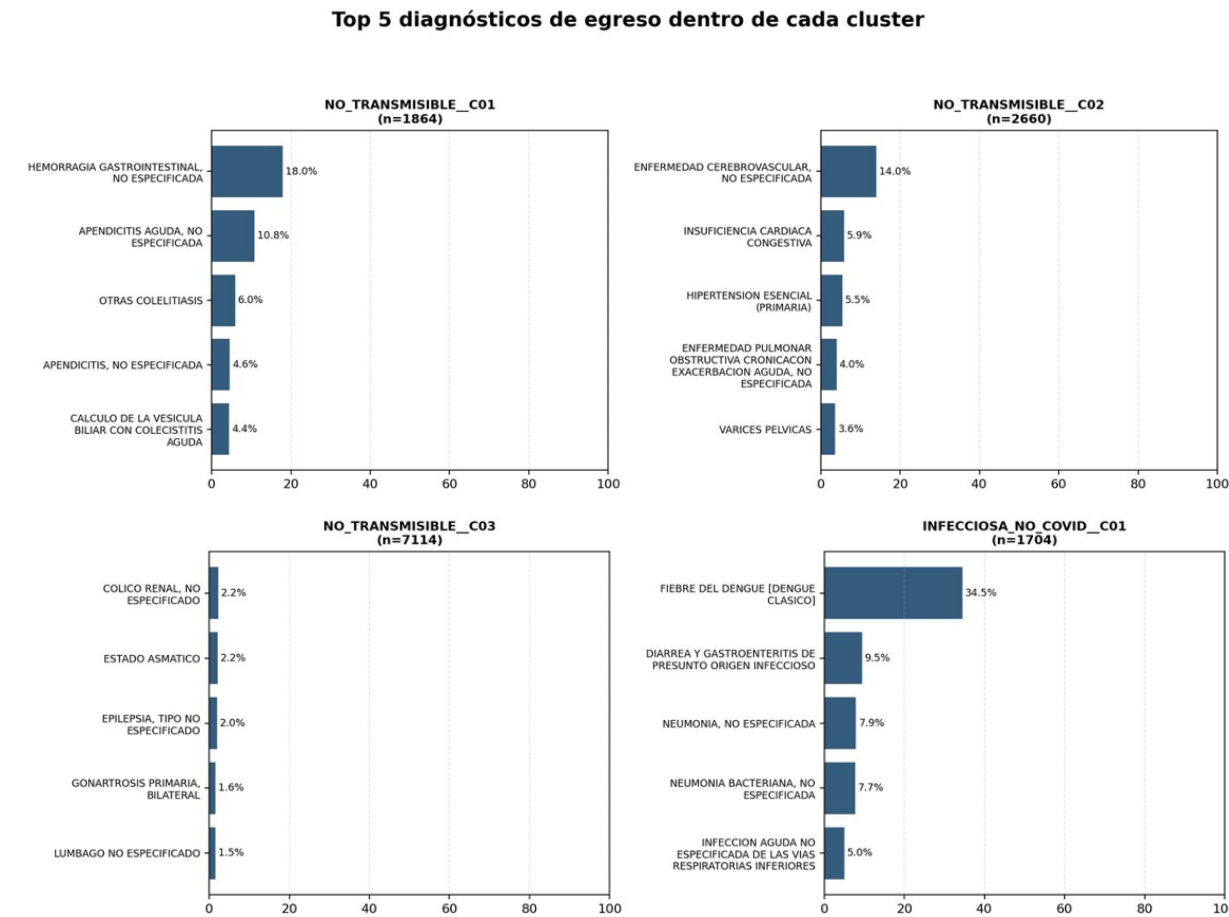
Distribución por sexo dentro de cada cluster

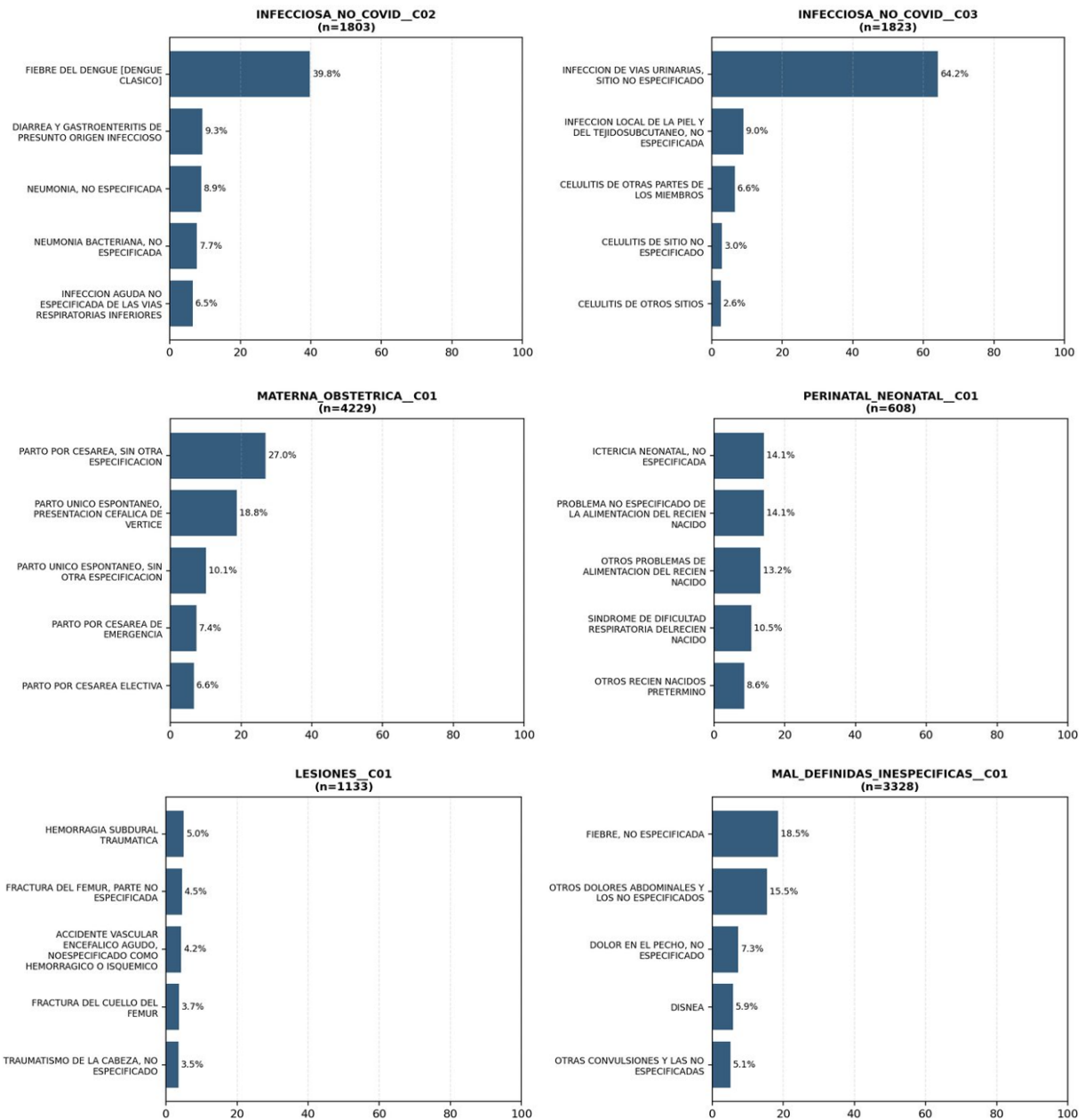




E.5.3.5 Diagnósticos de egreso más frecuentes. Este panel conserva la interpretación clínica de los grupos sin incorporar tablas extensas de diagnósticos en el cuerpo principal.

Figura 38
Diagnósticos de egreso más frecuentes dentro de cada grupo clínico final

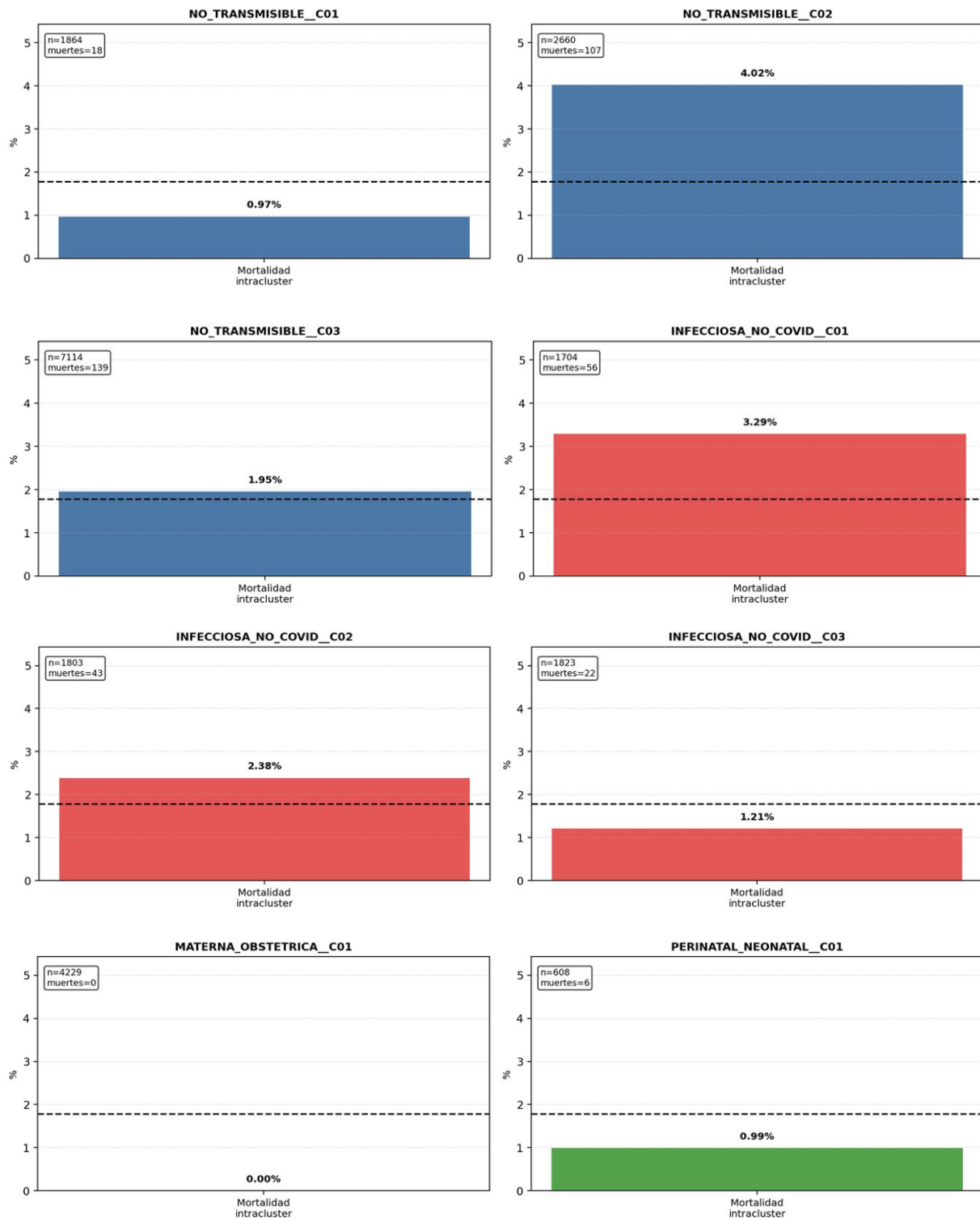


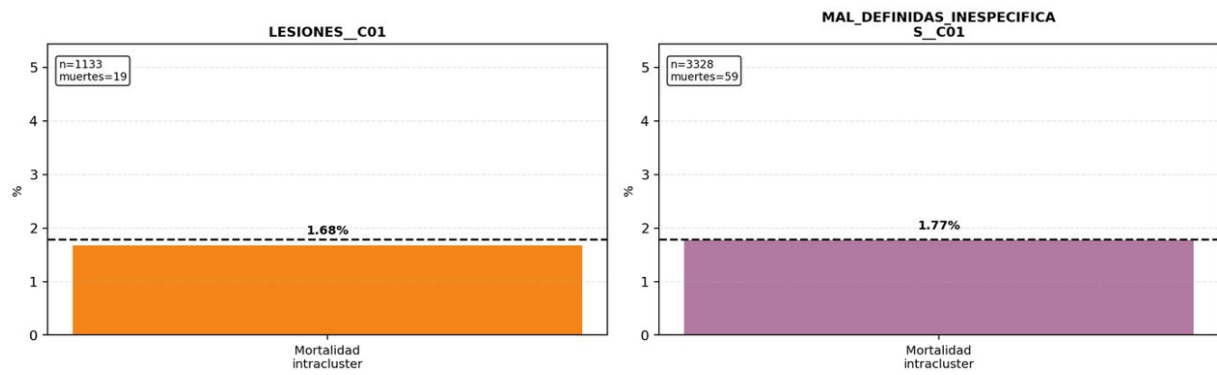


E.5.4 Paneles descriptivos de mortalidad

Los paneles de mortalidad complementan visualmente las tablas de mortalidad. La mortalidad se calculó después del agrupamiento y no participó en la formación de los grupos.

E.5.4.1 Mortalidad intragrupo.

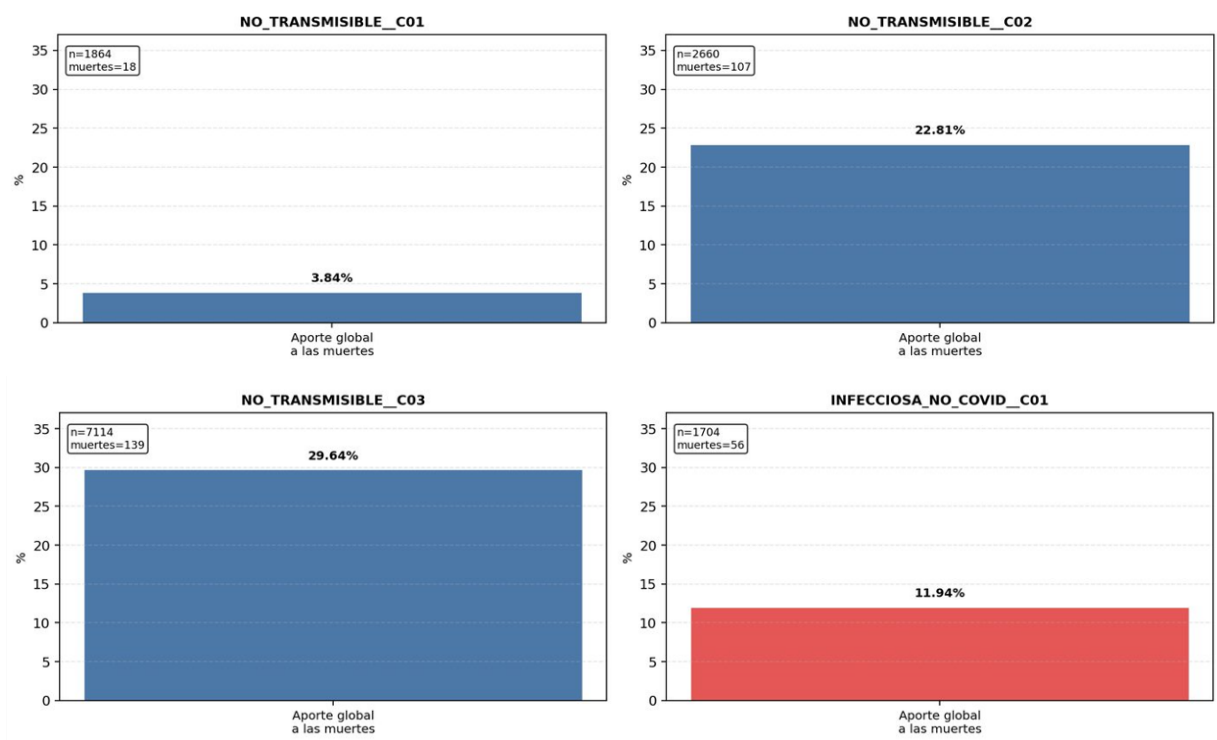
Figura 39*Mortalidad intragrupo por grupo clínico final***Panel de mortalidad intracluster**

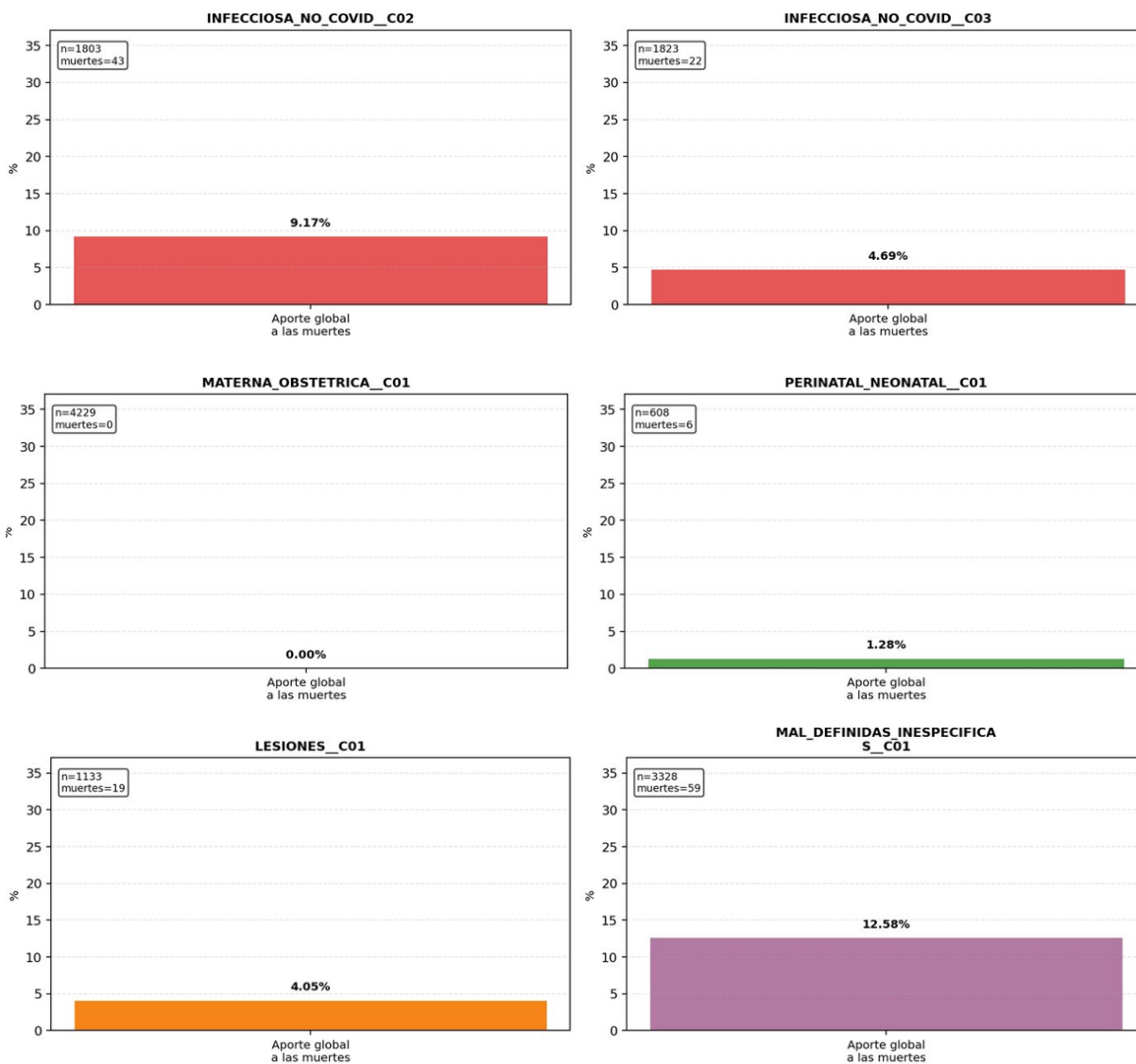


E.5.4.2 Aporte global a muertes.

Figura 40
Aporte global de cada grupo clínico final a las muertes

Panel de aporte global de cada cluster a las muertes



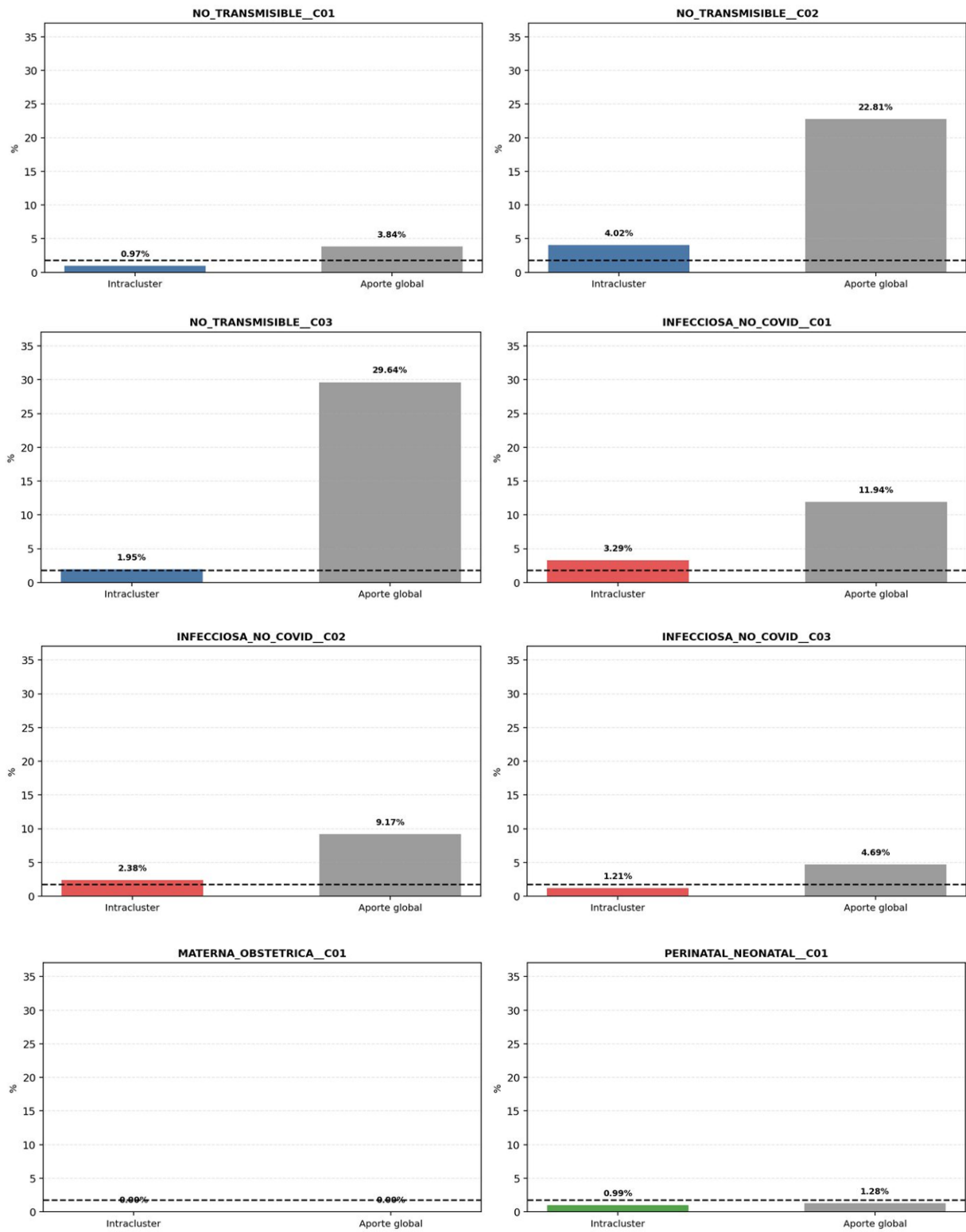


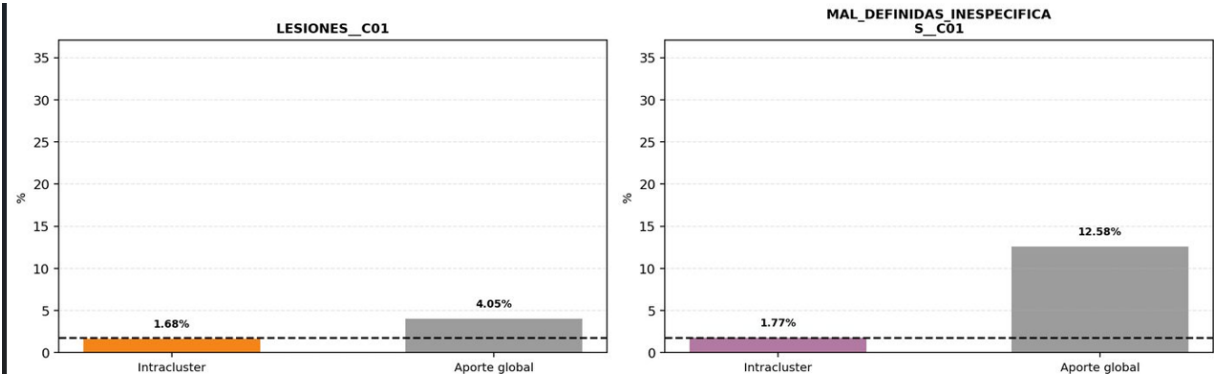
E.5.4.3 Comparación entre mortalidad local y aporte global.

Figura 41

Comparación entre mortalidad intragrupo y aporte global a muertes

Panel comparativo: mortalidad intracluster vs aporte global





E.6 Calidad del dato y controles técnicos

E.6.1 Tendencia y visualizaciones de calidad del dato

Figura 42
Comparación de indicadores de calidad del dato entre 2021–2023 y 2024

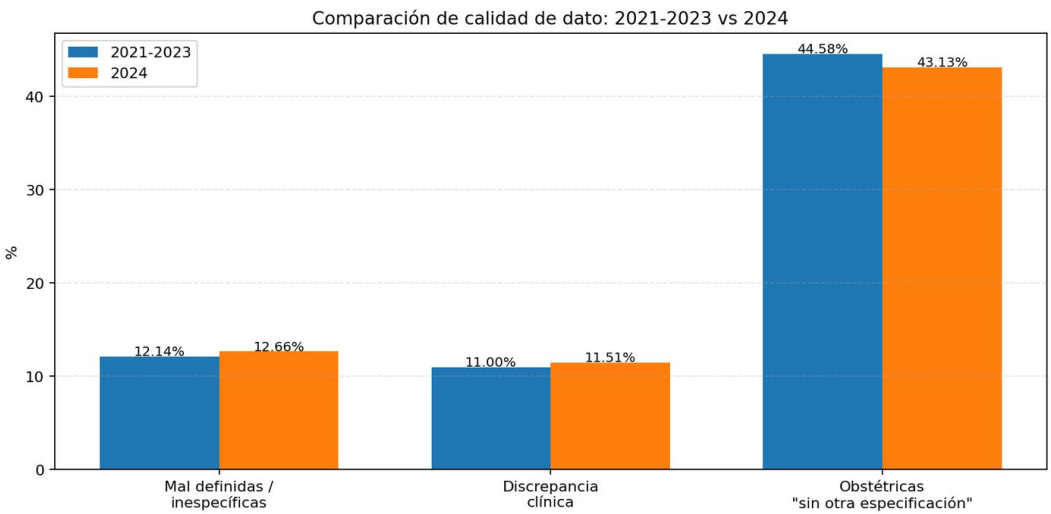


Figura 43
Tendencia anual de indicadores de calidad del dato

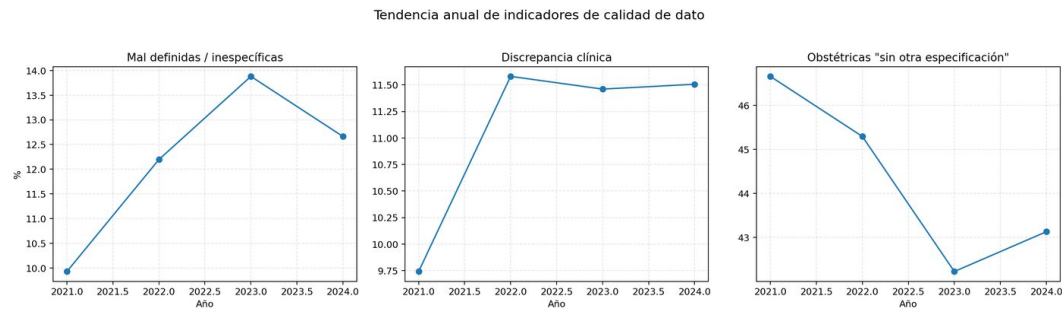


Tabla 81
Tendencia mensual de indicadores de calidad del dato durante 2024

Indicador	Pendiente pp/mes	r	p tendencia	Meses
Mal definidas o inespecíficas	-0,447	-0,869	< 0,001	12
Obstétricas sin otra especificación	-0,020	-0,095	0,7681	12

Nota. La tendencia se presenta como control de calidad temporal, no como criterio de selección de la cohorte. El indicador obstétrico aproxima especificidad diagnóstica: una menor proporción de registros sin otra especificación sugiere mayor detalle del diagnóstico consignado.

Estas figuras y la tabla mensual complementan el control de calidad del dato. El capítulo principal conserva la interpretación metodológica y este apéndice aloja los apoyos visuales y técnicos.

E.6.2 Faltantes, plausibilidad, valores atípicos y diagnósticos raros**Tabla 82***Datos faltantes en variables clave de la base final*

Variable	<i>n</i> faltantes	% faltantes
grupo_principal	0	0,00
macro_grupo_cluster	0	0,00
fall_bin	0	0,00
edad_num_modelo	0	0,00
grupo_edad_modelo	0	0,00
sexo_modelo	0	0,00
dx_codigo_modelo	0	0,00
dx_nombre_modelo	0	0,00
cie_3_modelo	0	0,00
servicio_modelo	0	0,00
uo_modelo	0	0,00
flag_discrepancia_clinica	0	0,00

Nota. Cada variable se evaluó sobre 26.266 registros finales; no se detectaron faltantes en las variables clave usadas para los análisis finales.

Tabla 83*Revisión de plausibilidad clínica básica*

Criterio	<i>n</i>	% cohorte
Diagnóstico materno-obstétrico en sexo masculino	2	0,008
Diagnóstico materno-obstétrico con edad < 10 o > 60	6	0,023
Diagnóstico neonatal con edad > 1 año	56	0,213

Nota. Estos conteos señalan registros para auditoría de calidad; no equivalen a errores clínicos confirmados.

Tabla 84*Resumen de valores atípicos por regla IQR*

Variable	n válidos	P25	Mediana	P75	Máx.	% outliers
edad	26.266	23	47	71	109	0,00
estancia	26.266	1,00	2,00	4,00	193	11,20
ingreso_total	26.266	887.964	1.734.110	4.677.675	509.569.621	13,42

Nota. Los outliers se reportan como control descriptivo; el análisis no los eliminó automáticamente.

Estancia e ingreso no formaron parte de las variables de agrupamiento.

Tabla 85*Resumen de diagnósticos raros por nombre diagnóstico*

Variable	Categorías	Frec. 1	Frec. 2	Episodios ≤2	% episodios
dx_nombre_modelo	2.355	983	382	1.747	6,65

Nota. Se conserva el resumen global; los listados Top 100 se excluyen por extensión y bajo valor interpretativo para el apéndice.

E.6.3 Unicidad de episodio e información no disponible

Tabla 86*Verificación de unicidad del identificador de episodio en la base final*

Indicador	Valor
Filas de la base final	26.266
Episodios únicos	26.262
Filas duplicadas por episodio	8
Episodios con más de una fila	4

Nota. Bajo la definición operativa institucional, los reingresos asociados a la misma trayectoria asistencial se conservan dentro del mismo episodio y no generan nuevas unidades independientes de análisis.

Tabla 87*Información no disponible para análisis adicionales solicitados o sugeridos*

Tema	Estado	Evidencia
Identificador único de paciente	No encontrado	No hay columnas tipo documento, historia clínica, paciente o usuario.
Comorbilidades	No encontrado	No hay diagnósticos secundarios, Charlson, Elixhauser ni campos equivalentes detectados.
SMR	No calculable con base actual	Requiere tasas esperadas o población estándar externa por edad/sexo.
Validación clínica externa	No calculable con base actual	Requiere muestra validada por experto con clasificación de referencia.

Nota. Estos análisis no deben reportarse como ejecutados porque requieren información no disponible en la base actual.

Estos controles delimitan el alcance de los resultados. En particular, el informe no puede afirmar independencia longitudinal de pacientes ni ajustar por comorbilidades con la información disponible. Tampoco debe reportar razones de mortalidad estandarizadas ni validación clínica externa como si se hubieran ejecutado.

E.7 Correcciones avanzadas ejecutadas con la base disponible

Esta sección documenta análisis adicionales ejecutados después de la versión inicial del capítulo. Solo se incorporan resultados efectivamente generados por el paquete de correcciones avanzadas. Los análisis que no se ejecutaron por ausencia de objetos o columnas requeridas se reportan como trazabilidad metodológica, sin inferir resultados no disponibles.

E.7.1 Error de agregación de microgrupos a centroides**Tabla 88***Resumen del error de agregación microgrupos-centroides en el espacio SVD*

Macrogrupo	<i>n</i>	Microgrupos	Dist. media	P90	P95	Inercia/episodio
No transmisible	11.638	120	0,030	0,055	0,067	0,0013
Infecciosa no COVID	5.330	120	0,056	0,112	0,164	0,0098

Nota. La distancia corresponde a la separación euclidiana entre cada episodio y el centroide de su microcluster en el espacio SVD. Valores más bajos indican microgrupos más compactos.

E.7.2 Sensibilidad ante exclusión de diagnósticos CIE-3 raros**Tabla 89***Sensibilidad de la solución al excluir CIE-3 con frecuencia menor o igual a 2*

Macrogrupo	<i>n</i> original	<i>n</i> retenido	Excluidos	CIE-3 raros	ARI	NMI
No transmisible	11.638	11.339	299	212	0,950	0,922
Infecciosa no COVID	5.330	5.265	65	49	0,693	0,708

Nota. ARI y NMI se calcularon frente a la solución final en los episodios retenidos. La exclusión de diagnósticos raros tuvo impacto bajo en no transmisible y moderado en infecciosa no COVID.

E.7.3 Reproducibilidad de métricas internas

Tabla 90

Reproducibilidad de métricas internas en 30 corridas

Macrogrupo	Validación	Sil. media	Sil. mín.	Sil. máx.	CH media	DB media
No transmisible	Semilla	0,254	0,230	0,278	888,50	1,409
No transmisible	Remuestreo 80 %	0,251	0,233	0,277	869,47	1,378
Infecciosa no COVID	Semilla	0,275	0,232	0,308	1.419,98	1,550
Infecciosa no COVID	Remuestreo 80 %	0,273	0,235	0,308	1.201,40	1,542

Nota. Sil.: coeficiente de silueta; CH: Calinski-Harabasz; DB: Davies-Bouldin. La tabla resume variabilidad de métricas internas y complementa ARI/NMI.

E.7.4 Métodos alternativos exploratorios

Tabla 91

Comparación exploratoria con métodos alternativos sobre el espacio SVD

Macrogrupo	Método	<i>n</i>	ARI	NMI	Silueta
No transmisible	K-means directo	11.638	0,789	0,797	0,278
No transmisible	Ward directo en muestra de 5.000	5.000	0,655	0,656	0,233
No transmisible	DBSCAN exploratorio en muestra	5.000	0,067	0,378	0,671
Infecciosa no COVID	K-means directo	5.330	0,272	0,367	0,322
Infecciosa no COVID	Ward directo	5.330	0,679	0,686	0,307
Infecciosa no COVID	DBSCAN exploratorio en muestra	5.000	0,117	0,432	0,836

Nota. ARI y NMI comparan cada alternativa con la solución final. DBSCAN generó muchos grupos y ruido, por lo que se reporta como exploración metodológica, no como solución candidata principal.

E.7.5 Utilidad exploratoria frente a macrogrupos originales

Tabla 92

Comparación de modelos de mortalidad con macrogrupo frente a grupo clínico final

Modelo	AUC	Log-loss	Brier	AIC aprox.
Edad + sexo	0,712	0,0848	0,0174	1.344,36
Edad + sexo + macrogrupo	0,731	0,0837	0,0174	1.338,79
Edad + sexo + grupo final	0,746	0,0825	0,0173	1.328,34
Solo macrogrupo	0,592	0,0868	0,0175	1.382,31
Solo grupo final	0,652	0,0860	0,0175	1.377,25

Nota. Comparación exploratoria en conjunto de prueba del 30 %. La mejora del modelo con grupo final frente al macrogrupo respalda utilidad descriptiva adicional de la segmentación, sin convertirla en validación causal o predictiva definitiva.

E.8 Trazabilidad de figuras y resultados conservados

Las figuras que permanecen en el Capítulo 7 corresponden a elementos directamente asociados con la metodología de agrupamiento: sensibilidad de composición con y sin COVID-19, gráficos de varianza explicada por SVD, métricas internas para selección de k y estabilidad mediante ARI y NMI. Este apéndice conserva los paneles descriptivos por servicio, edad, unidad médica, sexo, diagnósticos y mortalidad porque cumplen una función interpretativa posterior.

El apéndice excluye los listados Top 100 de diagnósticos raros, el detalle fila a fila de estabilidad en 30 corridas y las tablas completas de categorías diagnósticas por grupo cuando no agregan interpretación adicional. Esas salidas pueden conservarse como respaldo computacional, pero no son necesarias en el documento principal ni en el apéndice defendible.

Apéndice F. Complementos del Capítulo 8

F.1 Exploración gráfica y comparación extendida de transformaciones

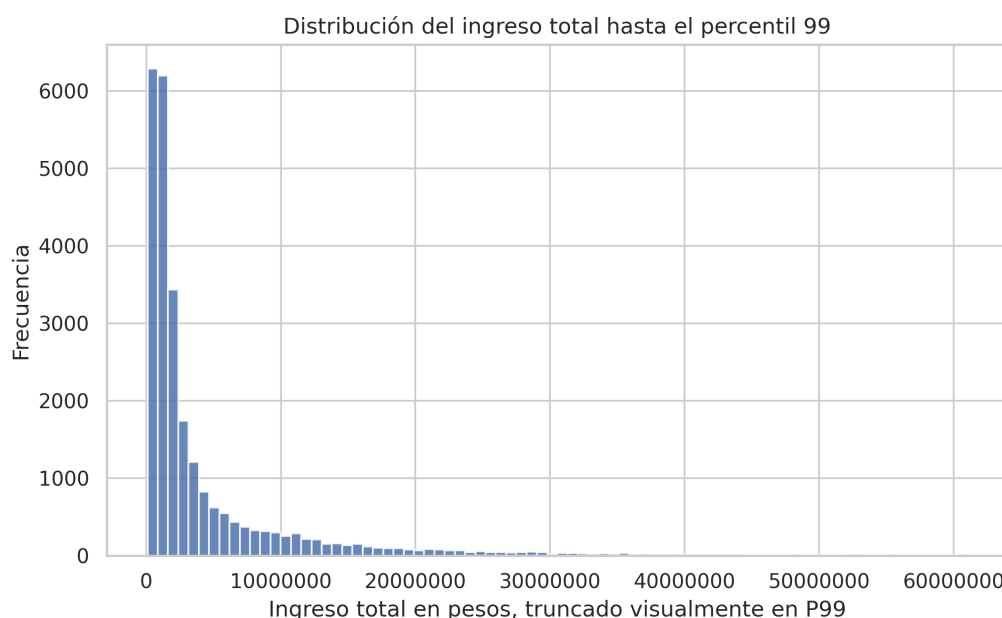
Este complemento documenta la evidencia que respalda la selección de la respuesta financiera transformada. El capítulo principal solo conserva las figuras indispensables; este apéndice reúne la comparación extendida para no sobrecargar la lectura central. La exploración consideró forma marginal, comportamiento residual, homocedasticidad, normalidad aproximada de residuos, ajuste del ANCOVA e interpretabilidad financiera.

F.2 Exploración complementaria de ingreso y transformación logarítmica

Esta sección amplía la exploración gráfica de la variable financiera principal. El Capítulo 8 presenta los histogramas esenciales de ingreso original y $\log(\text{Ingreso Total})$. Aquí se documentan gráficos adicionales para evaluar la forma de la distribución, la presencia de valores extremos y el efecto de la transformación logarítmica sobre la variable respuesta.

Figura 44

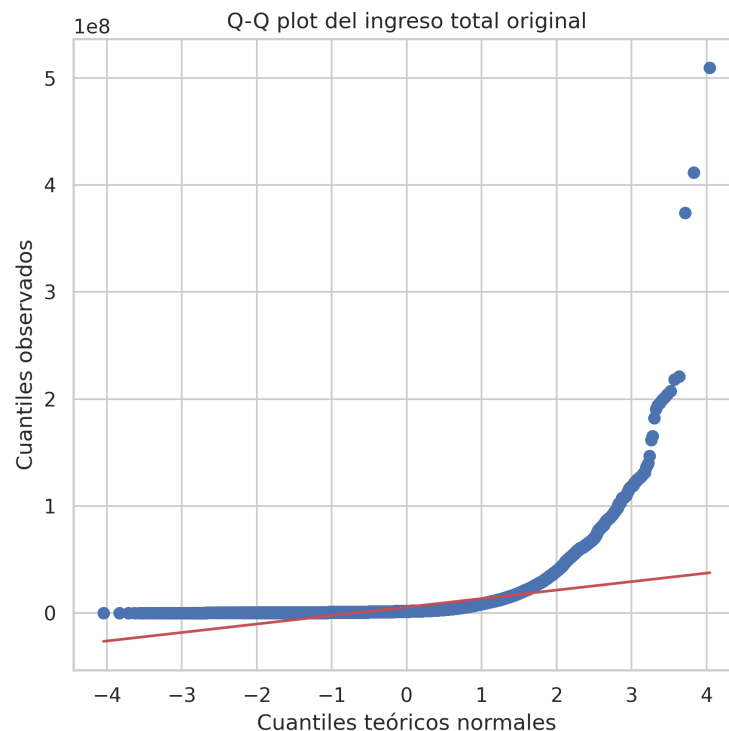
Histograma del ingreso total truncado visualmente en el percentil 99.



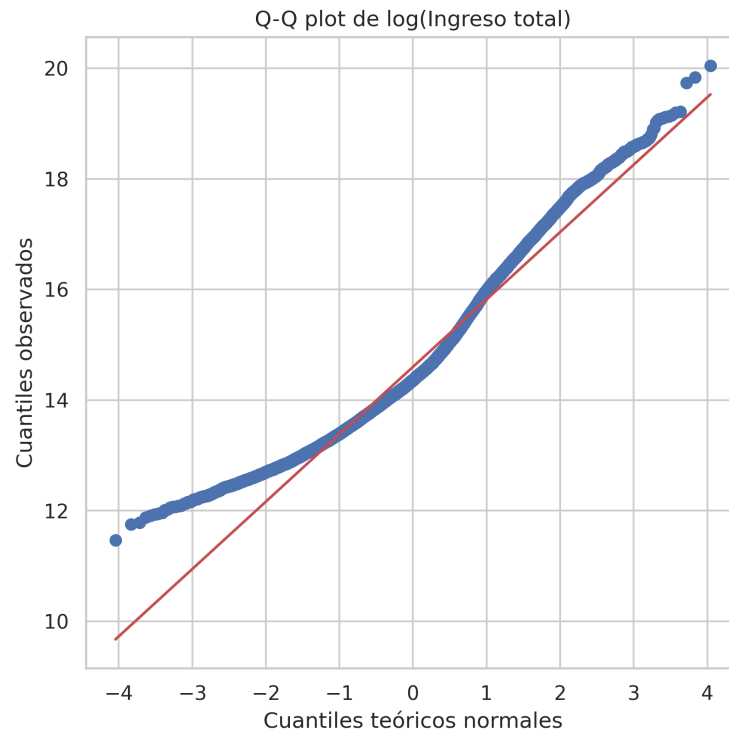
La Figura 44 muestra el cuerpo central de la distribución del ingreso total después de truncar visualmente el eje horizontal en el percentil 99. Este recorte no modifica los datos usados en los modelos; solo mejora la lectura gráfica de la mayoría de episodios, que en el histograma completo quedan comprimidos por la presencia de ingresos extremos. La figura confirma que, incluso al excluir visualmente el 1 % superior, la distribución conserva asimetría positiva.

Figura 45

Q-Q plot del ingreso total original.



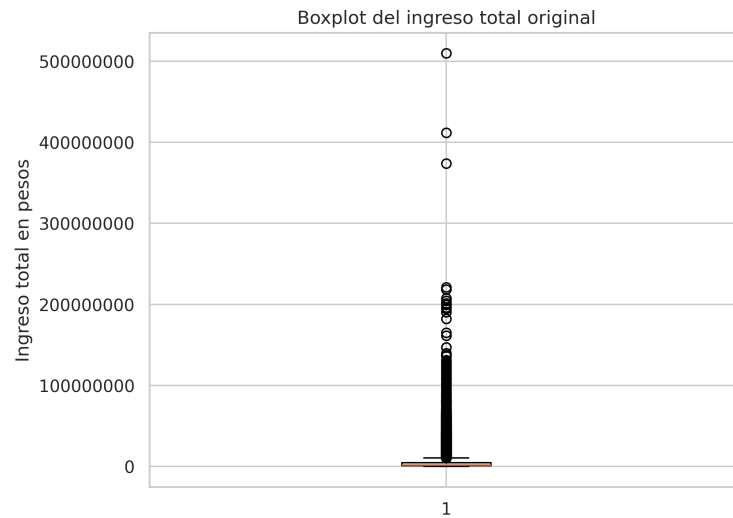
El Q-Q plot del ingreso total original muestra desviaciones importantes respecto a la recta de referencia normal, especialmente en la cola superior. Esta evidencia indica que la escala monetaria original no cumple una aproximación razonable a normalidad y que los valores altos ejercen una influencia marcada sobre la distribución. Por tanto, modelar directamente el ingreso original mediante un modelo lineal clásico habría concentrado la inferencia en los episodios de mayor valor.

Figura 46*Q-Q plot de $\log(\text{Ingreso Total})$.*

El Q-Q plot de $\log(\text{Ingreso Total})$ muestra una aproximación visual más cercana a la recta de referencia en la región central de la distribución. Sin embargo, todavía se observan desviaciones en las colas. Este resultado justifica dos decisiones metodológicas: primero, usar una transformación logarítmica para reducir la asimetría; segundo, mantener diagnósticos posteriores del modelo y errores estándar robustos, debido a que la transformación no elimina completamente las irregularidades distributivas.

Figura 47

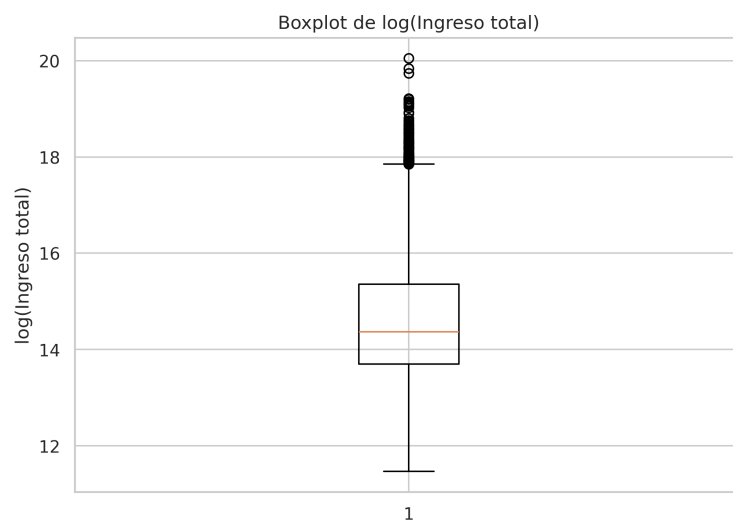
Boxplot global del ingreso total original.



La Figura 47 evidencia la presencia de numerosos valores alejados en la escala monetaria original. Estos valores no se eliminaron automáticamente, porque pueden representar episodios hospitalarios reales de alta complejidad, uso intensivo de recursos, tecnologías costosas o rutas clínicas específicas. El análisis los trató como observaciones potencialmente influyentes y los evaluó posteriormente mediante diagnóstico de residuos y distancia de Cook.

Figura 48

Boxplot global de log(Ingreso Total).



La Figura 48 muestra que la transformación logarítmica reduce la escala relativa de los valores

extremos y permite observar mejor la dispersión central de la variable financiera. Aun así, persisten observaciones alejadas en la escala transformada. Este comportamiento resulta esperable en datos financieros hospitalarios y respalda la decisión de complementar la transformación con escalamiento robusto por mediana e IQR.

En conjunto, los histogramas, Q-Q plots y boxplots documentan que la elección de una escala logarítmica robusta no se basó únicamente en valores de asimetría y curtosis. La decisión se apoyó también en evidencia visual sobre la forma de la distribución antes y después de la transformación. Estos resultados atienden la necesidad de mostrar explícitamente la distribución de la variable financiera antes de transformarla y de justificar la transformación empleada en el modelo principal.

Figura 49

Distribución de la respuesta financiera transformada por grupo clínico.

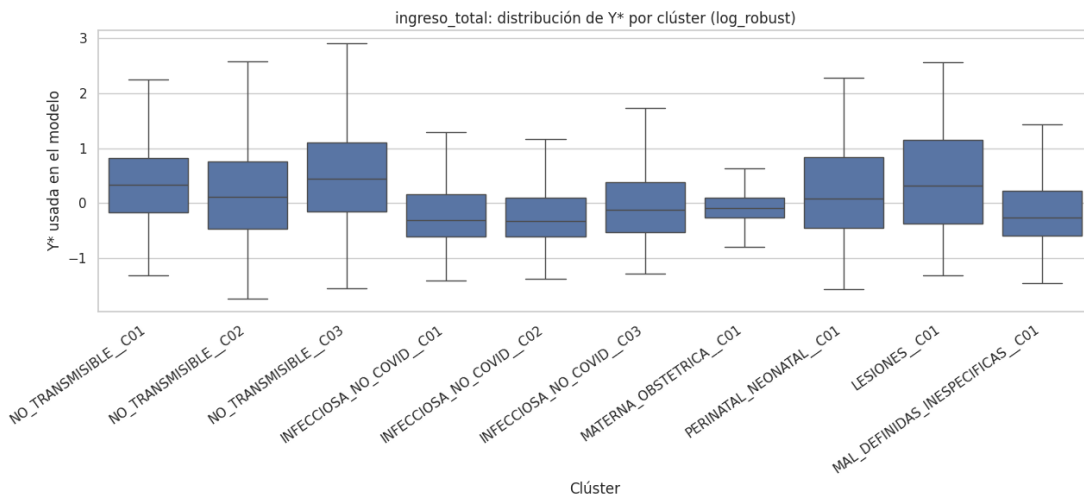
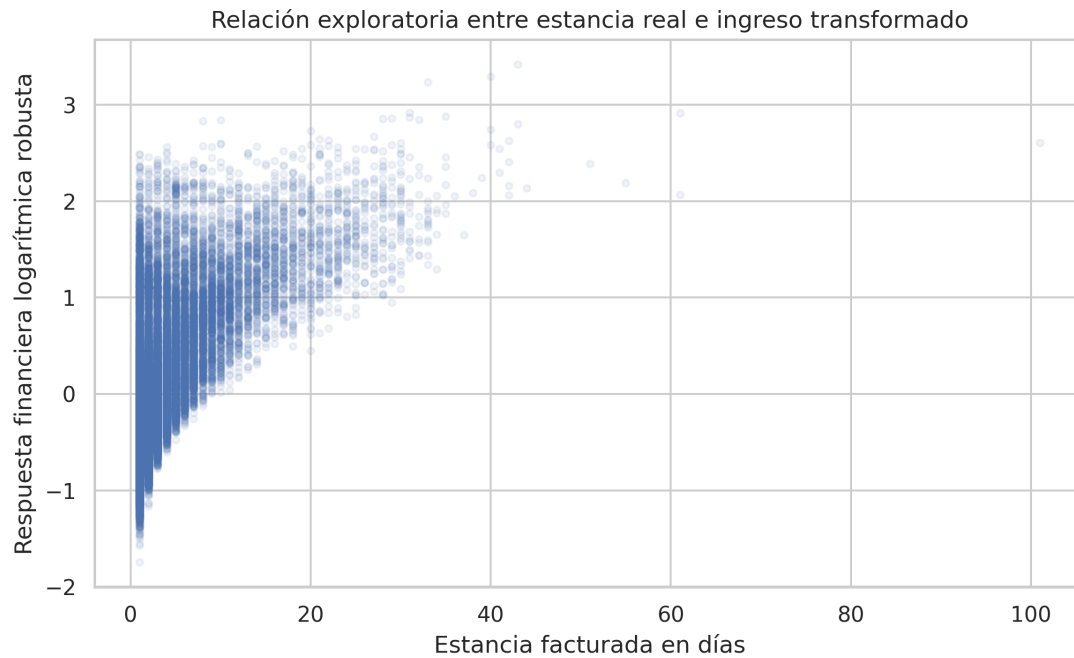


Figura 50

Relación global entre estancia real e ingreso transformado.

**Figura 51**

Relación entre estancia real e ingreso transformado por grupo clínico.

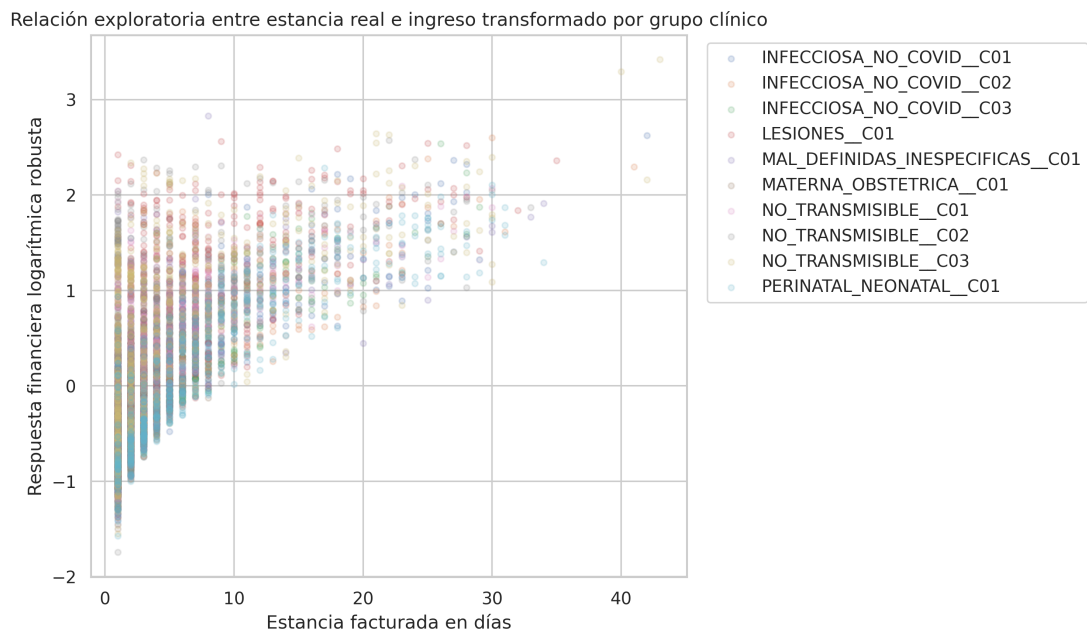


Tabla 93*Diagnóstico comparativo de transformaciones de la respuesta financiera*

Transformación	Asim. Y	Curt. Y	Asim. resid.	Curt. resid.	Normalidad resid. p	BP p	$R^2_{aj.}$
Ingreso original estandarizado	9,458	203,619	10,139	263,652	0,000	$7,04 \times 10^{-196}$	0,355
Logaritmo	0,692	0,028	0,527	1,542	0,000	0,000	0,481
Logaritmo robusto	0,692	0,028	0,527	1,542	0,000	0,000	0,481
Yeo–Johnson	0,063	-0,567	-0,152	0,836	$5,17 \times 10^{-108}$	0,000	0,435
Box–Cox	0,063	-0,567	-0,152	0,836	$5,16 \times 10^{-108}$	0,000	0,435
Winsorización + log	0,660	-0,167	0,496	1,570	0,000	0,000	0,478

Nota. BP corresponde a Breusch–Pagan. La prueba de normalidad residual corresponde a **normaltest**. En muestras grandes, valores p muy pequeños no implican por sí solos imposibilidad de modelamiento; por eso la decisión combinó forma marginal, residuos, ajuste e interpretabilidad. Yeo–Johnson y Box–Cox mejoraron la simetría marginal y residual, pero el logaritmo robusto conservó mejor interpretabilidad financiera y mayor ajuste que esas dos alternativas.

F.3 Diagnóstico del ANCOVA, influencia y supuestos operativos

El cuerpo del capítulo presenta el gráfico de residuos frente a valores ajustados porque resume la evidencia principal de heterocedasticidad. Este complemento no repite esa figura; en su lugar, agrega el Q–Q plot, la distancia de Cook y la tabla de episodios más influyentes. Esta organización evita duplicar figuras entre capítulo y apéndice.

Figura 52
Q-Q plot de residuos del ANCOVA con interacción.

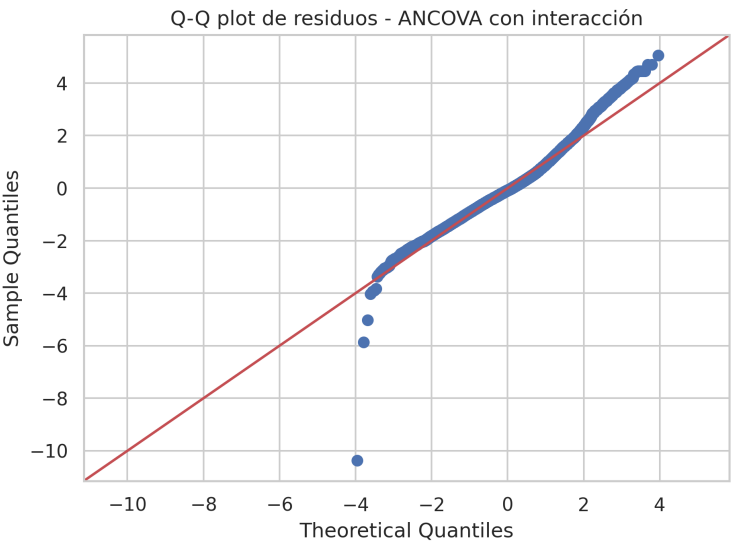


Figura 53
Distancia de Cook del ANCOVA con interacción.



Tabla 94
Episodios con mayor distancia de Cook en el ANCOVA con interacción

Grupo clínico	Ingreso	Estancia	Y*	Ajustado	Residuo	Cook
No transmisible C01	131.330.608	101	2,604	8,189	-5,585	1,456
No transmisible C02	218.395.627	61	2,910	6,067	-3,157	0,108

Grupo clínico	Ingreso	Estancia	Y^*	Ajustado	Residuo	Cook
Infecciosa no COVID C03	26.747.926	37	1,647	3,811	-2,164	0,025
Infecciosa no COVID C02	94.947.340	42	2,409	4,047	-1,638	0,022
Infecciosa no COVID C02	78.511.801	41	2,295	3,936	-1,641	0,021
Lesiones C01	16.105.058	33	1,341	3,434	-2,093	0,020
Infecciosa no COVID C01	136.137.482	42	2,626	4,136	-1,510	0,017
No transmisible C03	53.804.923	61	2,067	4,768	-2,701	0,016

La tabla muestra que los casos más influyentes combinan ingresos altos y estancias prolongadas. El análisis conservó estos episodios porque pueden representar observaciones reales de alta complejidad. Su presencia justifica la lectura robusta de la inferencia y la revisión de sensibilidad, no una eliminación automática.

F.4 Estimaciones marginales ajustadas completas y comparaciones post hoc

El capítulo principal reporta estimaciones ajustadas para los diez grupos. Este complemento extiende esa tabla con intervalos de confianza al 95 % para los mismos diez grupos y los tres niveles reales de estancia: 1, 2 y 5 días.

Tabla 95

Estimaciones marginales ajustadas completas con intervalos de confianza aproximados

Grupo clínico	Días	Predicción	IC95 % inf.	IC95 % sup.
Infecciosa no COVID C01	1	745.705	708.429	784.944
Infecciosa no COVID C01	2	900.111	858.949	943.246
Infecciosa no COVID C01	5	1.583.007	1.516.125	1.652.838
Infecciosa no COVID C02	1	725.342	690.150	762.329

Grupo clínico	Días	Predicción	IC95 % inf.	IC95 % sup.
Infecciosa no COVID C02	2	872.992	834.482	913.278
Infecciosa no COVID C02	5	1.521.987	1.458.458	1.588.283
Infecciosa no COVID C03	1	855.574	812.702	900.708
Infecciosa no COVID C03	2	1.040.325	993.065	1.089.835
Infecciosa no COVID C03	5	1.870.268	1.794.296	1.949.456
Lesiones C01	1	1.801.141	1.690.181	1.919.385
Lesiones C01	2	2.150.156	2.027.768	2.279.930
Lesiones C01	5	3.657.940	3.472.056	3.853.775
Mal definidas o inespecíficas C01	1	943.600	911.209	977.141
Mal definidas o inespecíficas C01	2	1.147.588	1.111.936	1.184.383
Mal definidas o inespecíficas C01	5	2.064.338	1.992.913	2.138.322
Materna-obstétrica C01	1	1.115.743	1.075.304	1.157.703
Materna-obstétrica C01	2	1.386.562	1.348.677	1.425.511
Materna-obstétrica C01	5	2.661.116	2.513.366	2.817.551
No transmisible C01	1	1.765.274	1.679.066	1.855.907
No transmisible C01	2	2.022.228	1.931.371	2.117.358
No transmisible C01	5	3.040.073	2.918.923	3.166.251
No transmisible C02	1	1.609.186	1.547.513	1.673.317
No transmisible C02	2	1.905.981	1.838.444	1.975.998
No transmisible C02	5	3.167.054	3.057.082	3.280.982
No transmisible C03	1	2.310.497	2.253.414	2.369.026
No transmisible C03	2	2.624.042	2.563.697	2.685.807
No transmisible C03	5	3.843.855	3.764.740	3.924.633
Perinatal-neonatal C01	1	836.161	757.610	922.856

Grupo clínico	Días	Predicción	IC95 % inf.	IC95 % sup.
Perinatal-neonatal C01	2	982.787	896.729	1.077.104
Perinatal-neonatal C01	5	1.595.763	1.479.352	1.721.335

F.5 Detalle financiero del grupo No transmisible C03

Las siguientes figuras y tablas documentan el análisis descriptivo del grupo de mayor peso financiero acumulado. El cuerpo del capítulo solo conserva la tabla de subgrupos para mantener una lectura sintética.

Figura 54

Ingreso acumulado por subgrupo clínico en No transmisible C03.

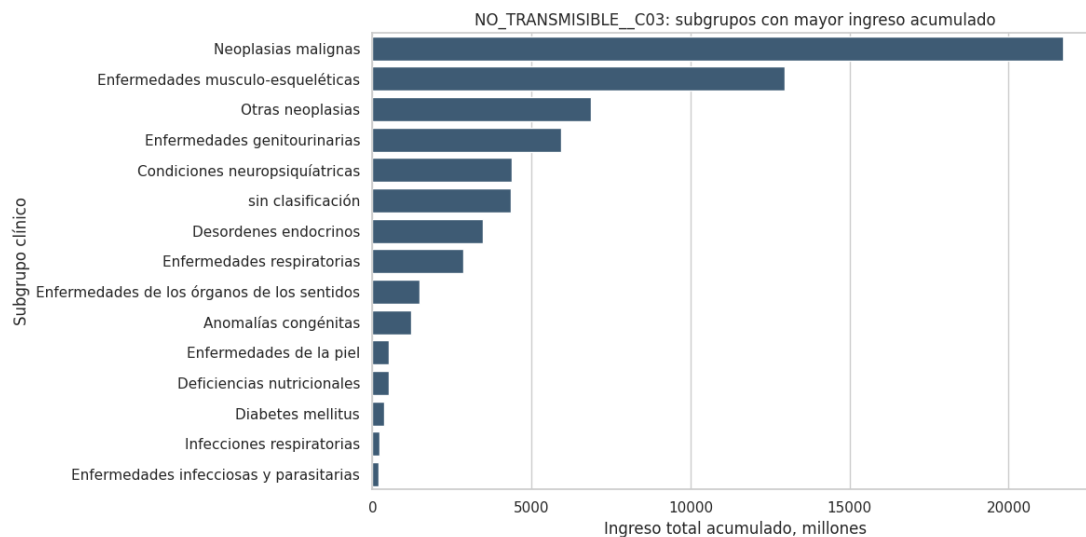


Figura 55
Diagnósticos con mayor ingreso acumulado en No transmisible C03.

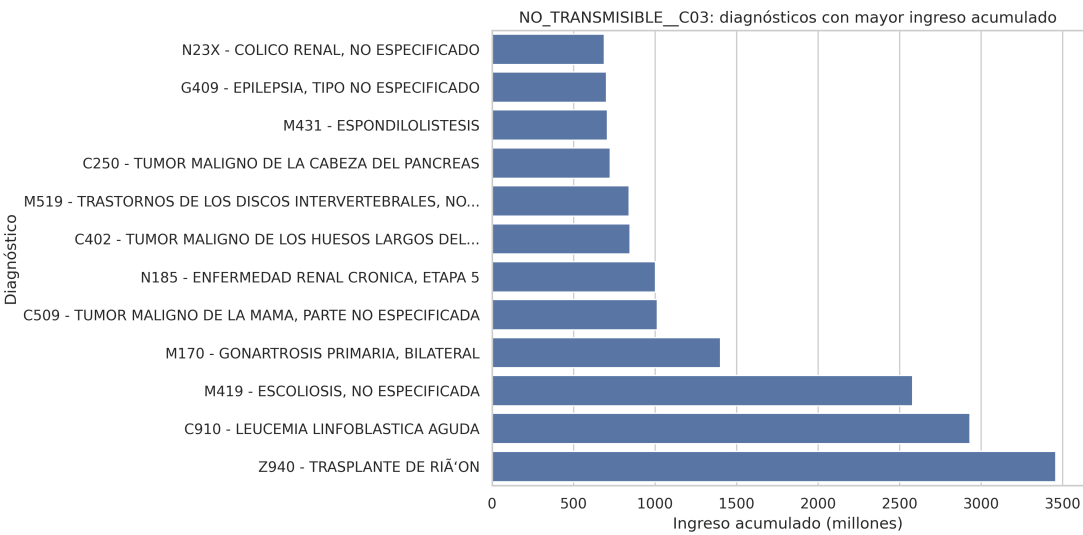


Figura 56
Diagnósticos con mayor ingreso mediano en No transmisible C03.

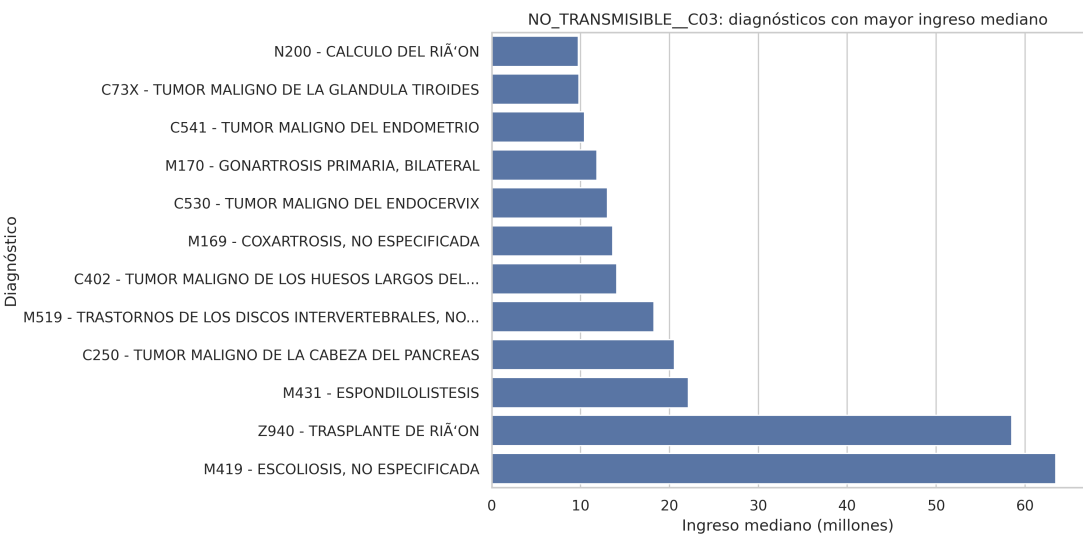


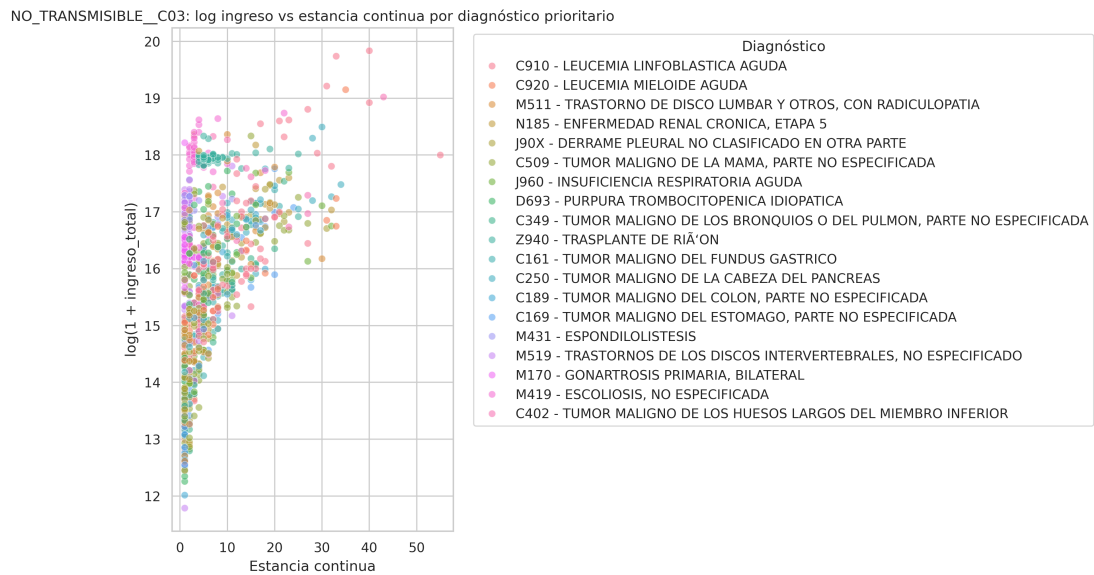
Figura 57*Relación entre estancia e ingreso en diagnósticos prioritarios de No transmisible C03.*

Tabla 96*Diagnósticos con mayor ingreso acumulado en No transmisible C03*

Diagnóstico	<i>n</i>	Ingreso acumulado	Mediana	Estancia med.
Z940 - Trasplante de riñón	81	3.458.502.785	58.491.492	7,0
C910 - Leucemia linfoblástica aguda	100	2.932.064.374	9.618.112	7,5
M419 - Escoliosis, no especificada	41	2.580.947.835	63.455.409	3,0
M170 - Gonartrosis primaria, bilateral	111	1.401.365.876	11.853.680	1,0
C509 - Tumor maligno de mama	100	1.013.536.512	7.589.139	4,0
N185 - Enfermedad renal crónica, etapa 5	75	1.002.657.723	6.028.371	6,0
C402 - Tumor maligno de huesos largos	30	846.537.935	14.070.252	7,0
M519 - Trastornos de discos intervertebrales	48	841.248.134	18.263.250	1,0
C250 - Tumor maligno de cabeza del páncreas	30	725.119.028	20.543.490	14,5
M431 - Espondilolistesis	35	708.811.027	22.117.960	2,0

F.6 Variabilidad financiera y variabilidad de estancia en diagnósticos prioritarios

Figura 58
Diagnósticos con mayor variabilidad financiera en No transmisible C03.

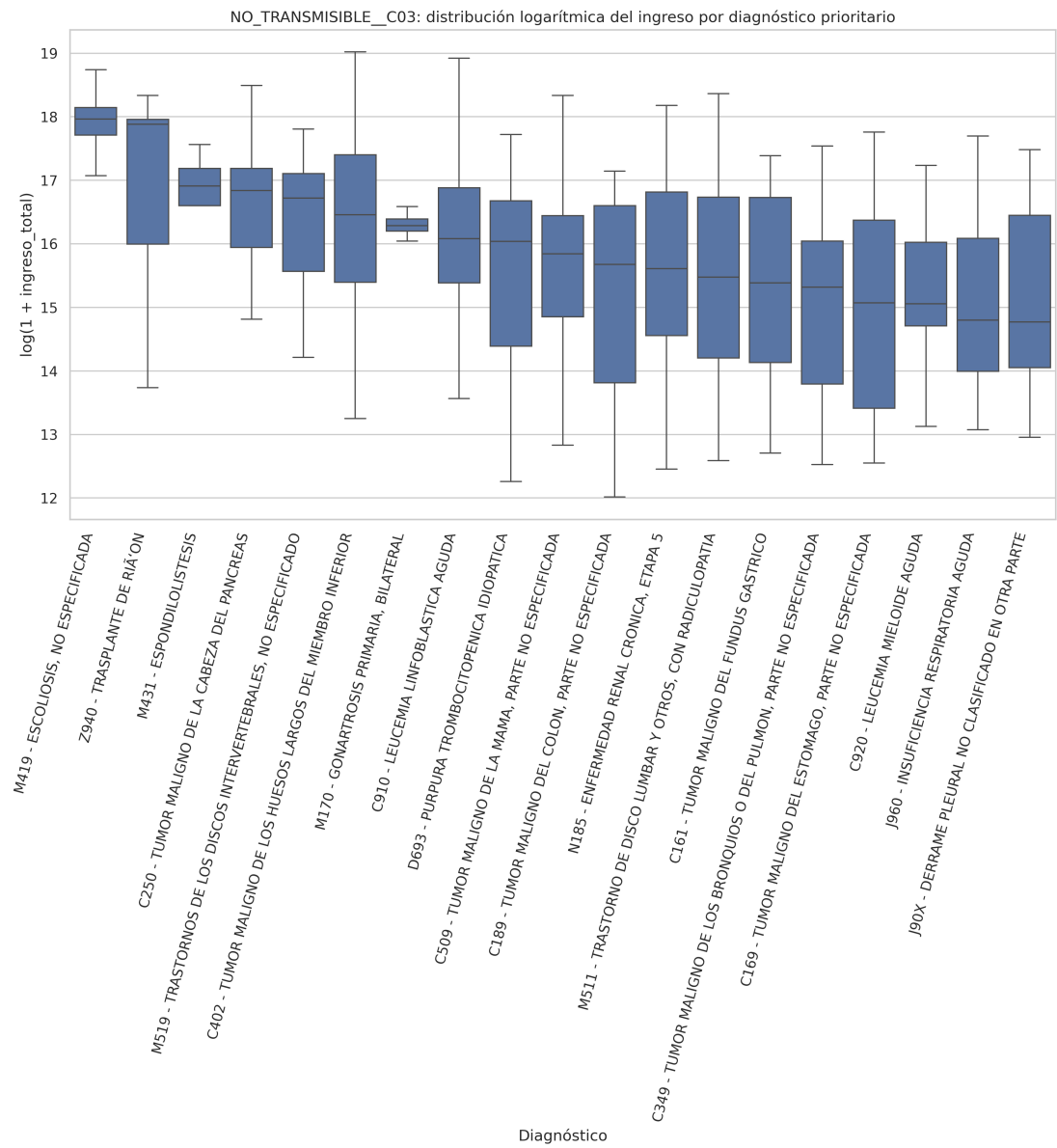


Figura 59
Diagnósticos con mayor variabilidad de estancia en No transmisible C03.

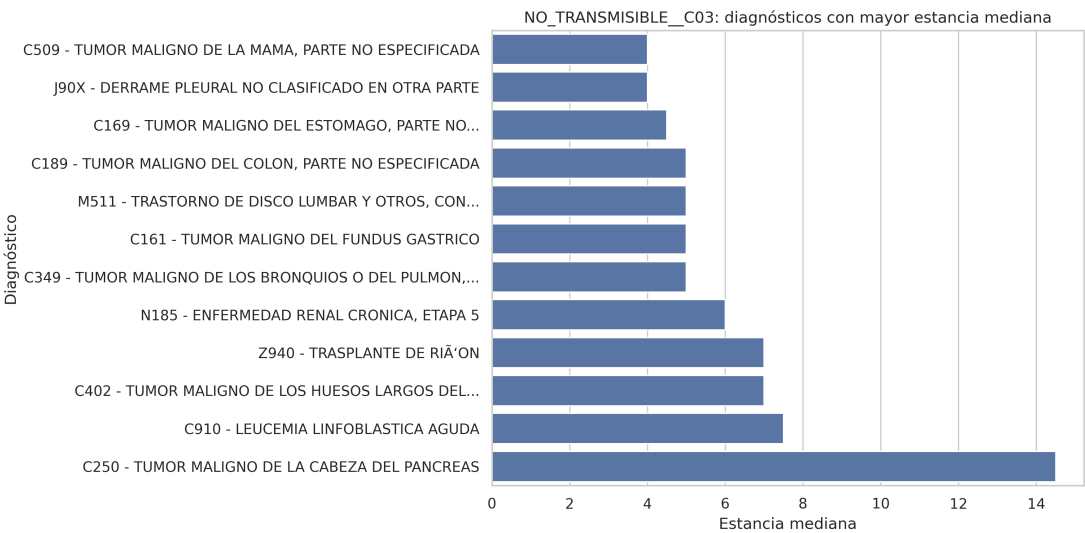


Tabla 97
Diagnósticos con mayor variabilidad financiera en No transmisible C03

Diagnóstico	n	Mediana	IQR ingreso	P90–P10
Z940 - Trasplante de riñón	81	58.491.492	54.112.241	60.951.488
C402 - Tumor maligno de huesos largos	30	14.070.252	31.269.672	62.964.520
M419 - Escoliosis, no especificada	41	63.455.409	26.773.265	79.823.636
M519 - Trastornos de discos intervertebrales	48	18.263.250	21.101.595	25.824.559
C250 - Tumor maligno de cabeza del páncreas	30	20.543.490	20.757.858	36.834.269
N185 - Enfermedad renal crónica, etapa 5	75	6.028.371	18.021.643	33.508.301
C910 - Leucemia linfoblástica aguda	100	9.618.112	16.709.027	63.451.829

F.7 Residuo financiero, variables descriptivas y residuo financiero positivo

Esta sección amplía los resultados del residuo financiero ajustado. Las figuras permiten explorar dónde se concentran residuos positivos por servicio, unidad médica, EPS y clase de movimiento de alta. Estos patrones no establecen causalidad; solo orientan revisión posterior.

Figura 60
Residuo financiero ajustado por servicio.

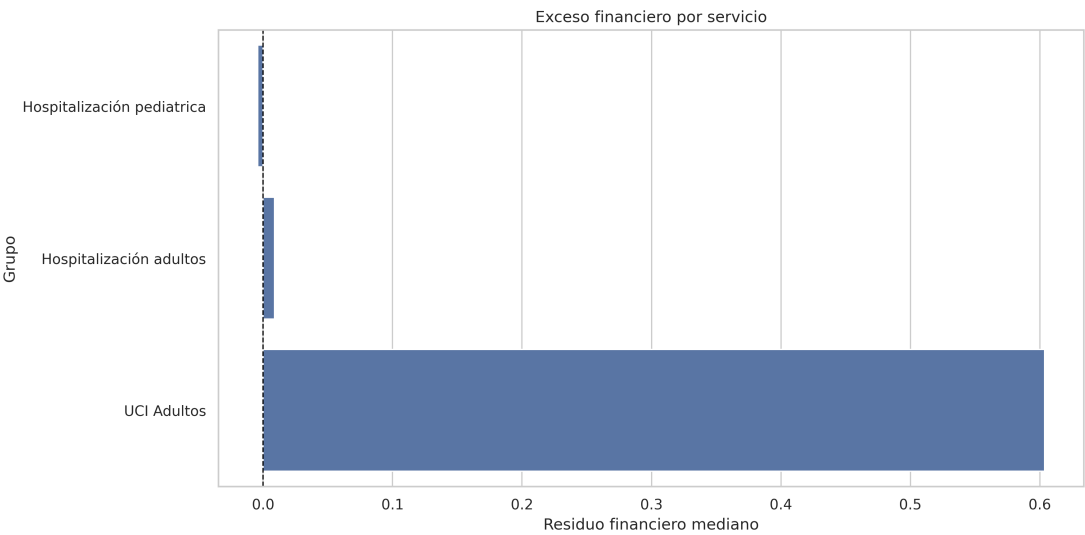


Figura 61
Residuo financiero ajustado por unidad médica.

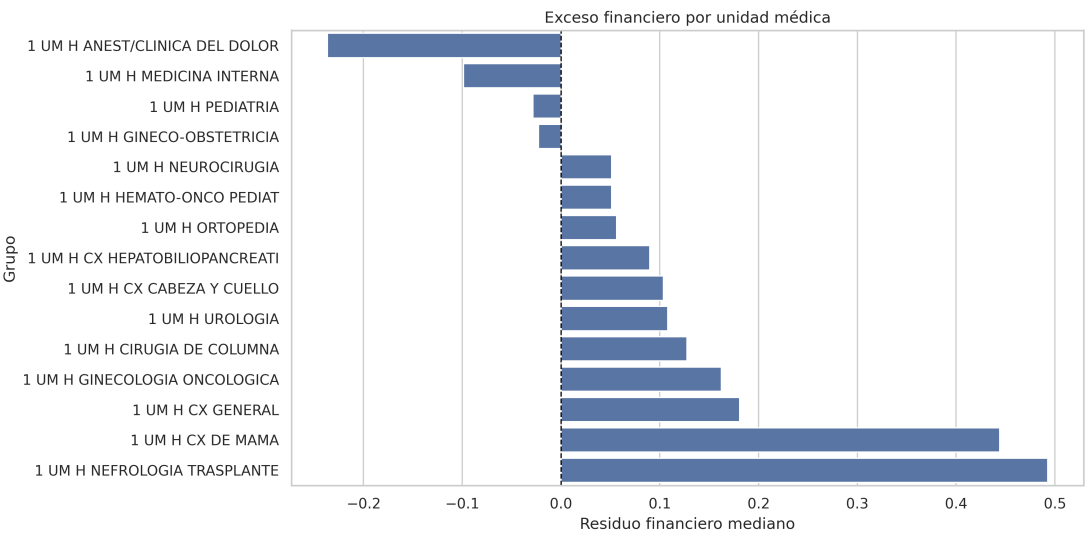


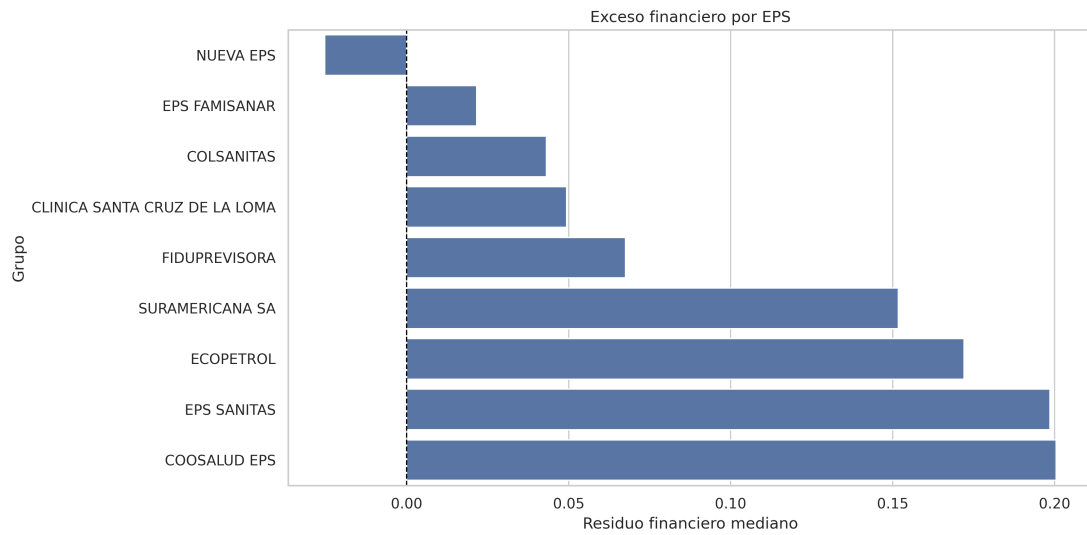
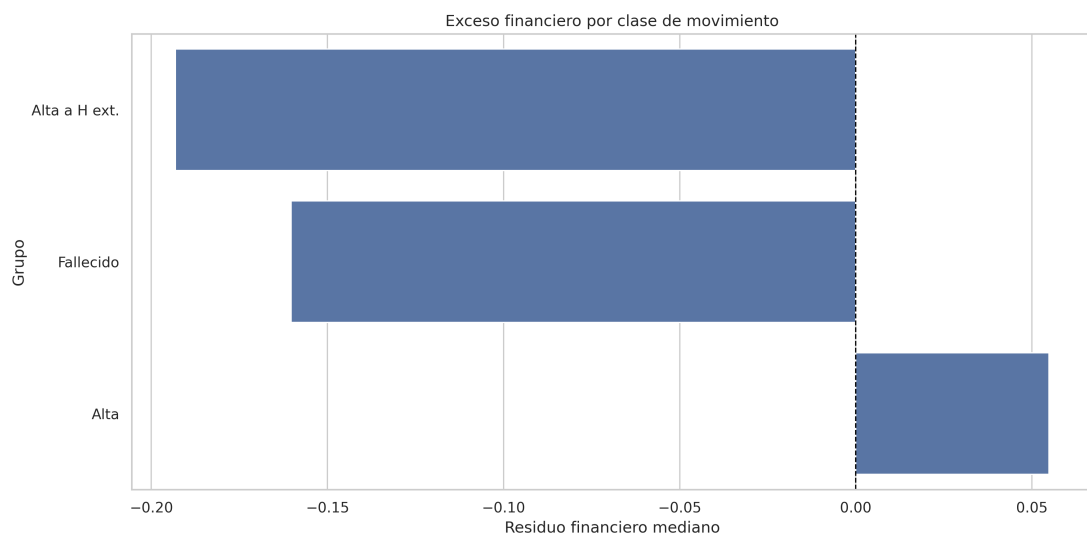
Figura 62*Residuo financiero ajustado por EPS.***Figura 63***Residuo financiero ajustado por clase de movimiento de alta.*

Tabla 98*Sensibilidad de la clasificación de residuo financiero positivo ante cambios de percentil*

Escenario	Categoría	<i>n</i>	Ingreso mediano	Residuo mediano
P75/P90	Alto	500	10.083.253	0,369
P75/P90	Muy alto	334	15.303.739	0,654
P70/P90	Alto	667	9.357.596	0,331
P70/P90	Muy alto	334	15.303.739	0,654
P80/P90	Alto	334	11.209.254	0,416
P80/P90	Muy alto	334	15.303.739	0,654
P75/P95	Alto	667	10.931.386	0,416
P75/P95	Muy alto	167	18.981.141	0,772
P80/P95	Alto	501	11.750.577	0,463
P80/P95	Muy alto	167	18.981.141	0,772

Nota. Los puntos de corte representan reglas exploratorias de priorización. La sensibilidad evalúa estabilidad interna; no equivale a validación externa.

Tabla 99*Concordancia de clasificaciones alternativas frente a P75/P90*

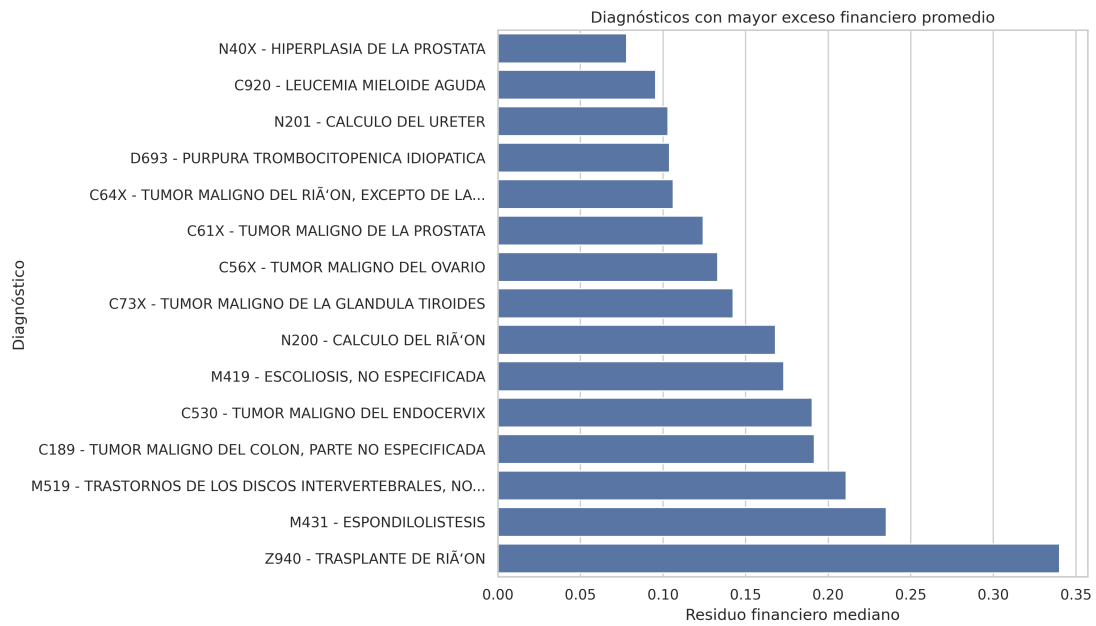
Comparación	ARI	NMI
P75/P90 frente a P70/P90	0,827	0,807
P75/P90 frente a P80/P90	0,820	0,794
P75/P90 frente a P75/P95	0,958	0,872
P75/P90 frente a P80/P95	0,787	0,668

La Tabla 98 muestra que la categoría de residuo financiero positivo muy alto conserva 334 episodios

cuando el punto de corte superior se mantiene en P90. Esto ocurre en los escenarios P70/P90, P75/P90 y P80/P90. En esos tres casos, el ingreso mediano del grupo muy alto permanece en 15.303.739 pesos y el residuo mediano en 0,654. Este resultado indica que la categoría más extrema no depende del punto de entrada usado para definir el grupo de exceso alto.

Cuando el punto de corte superior se eleva de P90 a P95, la categoría de residuo financiero positivo muy alto se vuelve más restrictiva: el número de episodios disminuye de 334 a 167, mientras que el ingreso mediano aumenta a 18.981.141 pesos y el residuo mediano a 0,772. Este comportamiento resulta coherente con una regla de priorización por cuantiles, porque los cortes más exigentes seleccionan menos episodios, pero concentran casos con mayor desviación financiera ajustada.

La Tabla 99 complementa esta lectura mediante medidas de concordancia. La regla principal P75/P90 presentó concordancia moderada-alta frente a los escenarios alternativos, con ARI entre 0,787 y 0,958 y NMI entre 0,668 y 0,872. Estos resultados respaldan estabilidad interna de la clasificación ante cambios razonables de percentil. Sin embargo, esta evidencia no constituye validación externa. Para validar institucionalmente la clasificación sería necesario contrastarla con información independiente, como auditorías reales, glosas, costos confirmados, consumo de insumos, procedimientos, contratos o revisión experta de casos.

Figura 64*Diagnósticos asociados a residuo financiero positivo muy alto.*

F.8 Control de reproducibilidad y datos faltantes para validación externa**Tabla 100***Información externa necesaria para validación futura*

Necesidad	Información requerida	Uso en una fase posterior
Independencia entre pacientes	Identificador anonimizado de paciente o historia clínica.	Estimar episodios repetidos y considerar modelos con efectos aleatorios o errores agrupados.
Validación de residuo financiero positivo	Auditorías reales, glosas, costos confirmados, insumos, procedimientos y contratos.	Evaluar si los residuos altos predicen hallazgos operativos o financieros reales.
Rentabilidad	Costos directos, costos indirectos, tarifas y margen por episodio.	Diferenciar ingreso alto de rentabilidad alta.
Causalidad operativa	Variables simultáneas de servicio, unidad, EPS, procedimiento y severidad.	Ajustar modelos multivariados más completos y evitar atribuciones descriptivas simples.