Policy brief

Lago de datos ÁGORA, integración de datos de salud a través de soluciones en Big Data

Andrés Moreno[1], Jennifer Murillo[2], Daniel Bonilla[2], Johan Manuel Calderón Rodríguez[2], Jenny Marcela Pinilla[2] ,Juan Guillermo Torres[3], Jaime A. Pavlich-Mariscal[1], Zulma M. Cucunubá[2]

Mensajes clave

- Los lagos de datos permiten integrar información dispersa y mejorar la toma de decisiones en salud pública.
- La experiencia del lago de datos ÁGORA demuestra que es posible construir soluciones técnicas con datos nacionales.
- Persisten retos críticos en gobernanza, sostenibilidad y acceso seguro a los datos.
- Es urgente institucionalizar estas herramientas de grandes datos para responder a futuras emergencias sanitarias.

Sobre el estudio ★

El lago de datos ÁGORA, desarrollado sobre infraestructura de alto rendimiento (Hadoop y Spark) en el clúster HPC-ZINE de la Pontificia Universidad Javeriana, integra más de 4.300 millones de registros de salud pública de Colombia (2009–2023), permitiendo análisis avanzados gracias a su estructura por zonas de datos y a la interoperabilidad segura mediante identificadores anonimizados. Su potencial se evidencia en cruces de datos clínicos y epidemiológicos con alta concordancia entre fuentes como RIPS, SegCovid y SIVIGILA, lo que permite estudios detallados sobre vacunación, mortalidad, trayectorias clínicas y desigualdades sanitarias. Un caso aplicado demostró su capacidad para clasificar causas de muerte y revelar brechas estructurales en salud. No obstante, enfrenta retos estructurales, normativos, técnicos y humanos, por lo que se recomienda fortalecer la gobernanza de datos, asegurar sostenibilidad e inversión, fomentar la colaboración entre entidades y consolidar su rol en la preparación ante futuras emergencias sanitarias.

1 Departamento de Ingeniería de Sistemas, Facultad de Ingeniería, Pontificia Universidad Javeriana, Bogotá, Colombia.
2 Instituto de Salud Pública, Pontificia Universidad Javeriana, Bogotá, Colombia
3 HPC-ZINE, Laboratorios Facultad de Ingeniería. Facultad de Ingeniería, Pontificia Universidad Javeriana, Bogotá, Colombia.

Problema

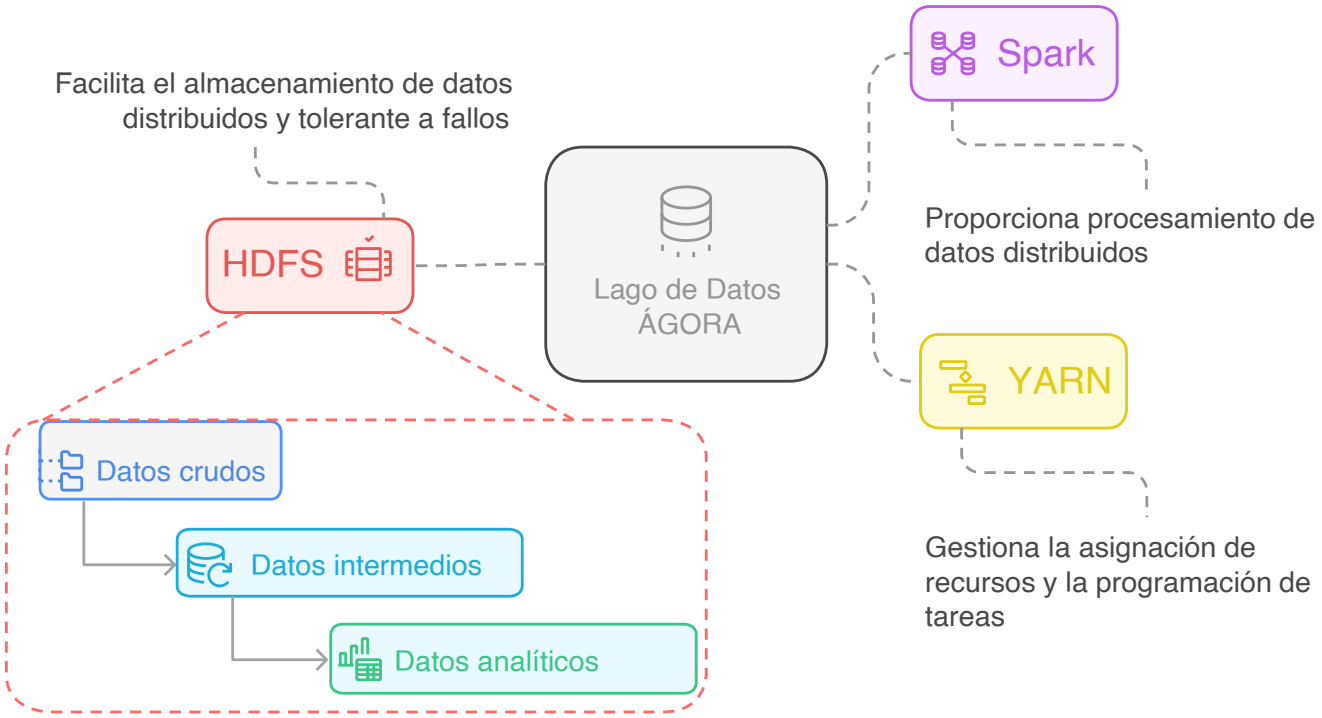
En Colombia, la información en salud pública ha estado distribuida entre diferentes fuentes y entidades, lo que ha limitado la capacidad del Estado para tomar decisiones oportunas, basadas en evidencia en tiempo real y adaptadas a las necesidades territoriales. Durante la pandemia por COVID-19, aunque existían múltiples bases de datos valiosas, como estadísticas vitales (defunciones y nacimientos), RIPS (Registros Individuales de Prestación de Servicios), SIVIGILA, SEG-COVID y PAIWEB, la ausencia de sistemas interoperables y accesibles que integren estas fuentes redujo la capacidad de anticipar escenarios, asignar recursos equitativamente y monitorear la efectividad de las intervenciones. Por ende, la creación de un sistema robusto y sostenible para integrar y analizar esta información es clave para fortalecer la respuesta ante emergencias futuras y mejorar el monitoreo rutinario de la salud pública.

Resultados

Infraestructura del Lago de Datos ÁGORA: integración, seguridad y análisis

El lago de datos ÁGORA se construyó soportado en una infraestructura de cómputo de alto rendimiento basada en Hadoop y Spark, tecnologías que permiten el procesamiento distribuido de grandes volúmenes de información. Está estructurado en tres zonas: datos crudos (sin procesar), datos intermedios (verificados y estandarizados) y datos analíticos (listas para análisis). Esta infraestructura permite realizar cálculos sobre grandes volúmenes de información tal y como agregaciones y cruces entre fuentes de manera eficiente. La interoperabilidad entre fuentes se logra mediante identificadores anonimizados no reversibles, lo que facilita cruces entre bases masivas. Todo el sistema opera en el clúster HPC-ZINE de la Pontificia Universidad Javeriana (Figura 1).

Figura 1: Arquitectura del Lago de Datos ÁGORA



En él Lago ÁGORA se consolidan los registros de salud pública de toda Colombia entre los años 2009 a 2023, correspondientes a bases de datos compartidas por el Ministerio de Salud y Protección Social (Estadísticas vitales, RIPS, SIVIGILA, SegCovid y vacunación COVID-19) junto con datos públicos abiertos obtenidos de diferentes entidades gubernamentales, con un volumen total de 4381 millones de registros individuales (Tabla 1).

Tabla 1: Volúmenes de información registrados en las fuentes más grandes presentes en el lago de datos

Base	Período	# registros (en millones de registros)
Defunciones	2014-01 a 2023-12	2,49
Nacimientos	2014-01 a 2023-12	5,90
RIPS	2009-01 a 2022-12	4.262,86
SegCovid	2020-03 a 2023-04	18,43
SIVIGILA (84 eventos de salud pública)	2009-01 a 2023-12	9,51
SIVIGILA-IRA por virus nuevo (ficha 346, que correspondió a COVID-19)	2020-03 a 2023-12	6,24
Vacunación	2021-01 a 2023-12	81,58
Total		4387,1

Cruces seguros y clasificación clínica: El potencial de los datos integrados

La Tabla 2 muestra los niveles de coincidencia entre registros individuales de diversas fuentes de información en salud en Colombia, entre 2014 y 2024. Se evidencia que los RIPS funcionan como un nodo central de integración, al presentar coincidencias superiores al 89% con todas las demás fuentes y más del 96% con SegCovid y SIV-346, lo que resalta su valor para vincular datos individualizados.

También se observa una alta concordancia entre SegCovid y SIV-346 (97,8%), validando su uso conjunto para el seguimiento de casos. Un hallazgo relevante es la fuerte conexión entre RIPS y SIV (91,1%). Esta relación destaca la posibilidad de integrar información de notificación obligatoria (SIV) con registros clínicos detallados (RIPS), fortaleciendo el análisis epidemiológico en tiempo real.

Las coincidencias más bajas se presentan con las bases de defunciones (DEF) y nacimientos (NAC), lo cual no refleja necesariamente fallas en la integración, sino más bien el hecho de que muchas personas registradas no han fallecido ni nacieron durante el periodo analizado.

Se destacan además relaciones relevantes como la coincidencia del 72,8% entre los registros de nacimientos (NAC) y vacunación (VAC), lo cual abre oportunidades valiosas para el seguimiento de cohortes jóvenes. Esta conexión es especialmente útil para evaluar programas de vacunación introducidos en años recientes, facilitando la construcción de cohortes prospectivas con datos integrados desde el nacimiento. Asimismo, la alta coincidencia entre VAC y SIV-346 (87,2%) permite analizar con mayor precisión la cobertura y el impacto de la vacunación frente al COVID-19. Aunque actualmente el registro VAC se limita a PAIWEB 2.0 que nació solo a partir de las vacunas contra COVID-19, su potencial crecerá significativamente en la medida en que se integren otras vacunas, consolidando un sistema de datos individualizados clave para la vigilancia y evaluación de programas de inmunización a nivel nacional.

En conjunto, estos cruces demuestran el enorme potencial de la interoperabilidad de datos para la analítica avanzada en salud pública, tanto en contextos de emergencia como en la vigilancia y planificación sanitaria de forma rutinaria.

Tabla 2: Porcentaje de coincidencia de IDs únicos entre fuentes, respecto a los IDs únicos de la fuente en la fila. DEF: Defunciones. NAC: Nacimientos. RIPS: Registro Individual de Prestación de Servicios. SIV: SIVIGILA (otros eventos). SIV-IRAvn: SIVIGILA-IRA por virus nuevo (ficha 346). SEG: SegCovid-19. VAC: Vacunación.

	DEF	NAC	RIPS	SEG	SIV	SIV-346	VAC
DEF	100,0%	0,7%	91,7%	8,6%	10,9%	8,4%	23,9%
NAC	0,4%	100,0%	97,1%	11,9%	20,5%	12,7%	72,8%
RIPS	3,2%	6,4%	100,0%	7,5%	9,0%	8,0%	48,7%
SEG	3,9%	10,1%	97,1%	100,0%	11,9%	97,8%	87,3%
SIV	3,9%	13,5%	91,1%	9,2%	100,0%	9,4%	54,0%
SIV-346	3,6%	10,1%	96,6%	90,7%	11,3%	100,0%	87,2%
VAC	1,5%	8,7%	89,2%	12,3%	9,8%	13,2%	100,0%

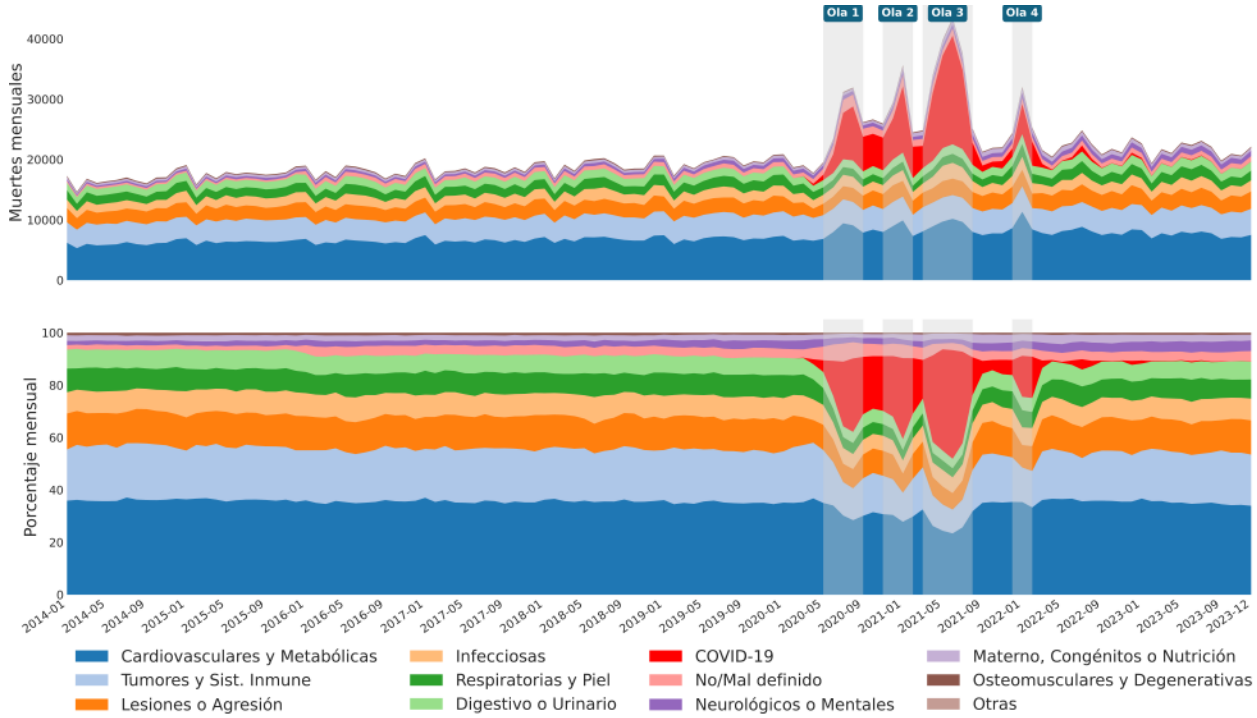
El alto cruce de fuentes integradas abre múltiples posibilidades de investigación en salud pública. Por ejemplo, permite evaluar la efectividad de la vacunación, estimar la letalidad ajustada por edad y comorbilidades, y analizar trayectorias clínicas a partir del uso de servicios de salud. También posibilita estudios sobre subregistro y brechas territoriales en la vigilancia, seguimiento de cohortes de nacimientos y vacunación infantil, así como análisis del impacto en la mortalidad materna e infantil. Además, se pueden estudiar factores de riesgo como enfermedades crónicas, evaluar la cobertura vacunal en poblaciones específicas, y desarrollar investigaciones metodológicas sobre interoperabilidad y calidad de los datos.

Caso de uso del lago ÁGORA: 10 años de causas de muerte en Colombia (2014 y 2024)

A través de este sistema, se realizó un caso de aplicación con la consolidación de más de 2.4 millones de registros de defunciones registradas entre el 1 de enero de 2014 y el 31 de diciembre de 2023, los cuales fueron enriquecidos mediante cruces con otras bases, utilizando identificadores anonimizados.

Este enfoque reveló patrones en la mortalidad por COVID-19, especialmente durante las olas epidémicas, y de otras causas como enfermedades cardiovasculares e infecciosas (Figura 2). También evidenció mayores cifras de causas mal definidas en regiones con menor capacidad diagnóstica, como la Amazonía y el Pacífico, lo que refuerza la utilidad del lago para identificar desigualdades estructurales en la infraestructura de información.

Figura 2: Principales causas de muerte durante el período enero de 2014 a diciembre de 2023 en Colombia (total y porcentaje) por categoría ÁGORA



Este proceso de clasificación de causas de muerte se realizó usando el sistema de clasificación propio del proyecto (agrupación ÁGORA), el cual permitió analizar de forma organizada más de 12.000 códigos CIE-10 y generar categorías clínicas relevantes para el contexto colombiano. A pesar de un subregistro leve frente al DANE (3,3 % en promedio), los análisis realizados muestran el potencial del lago de datos como herramienta estratégica para generar evidencia en salud pública.

Lago de datos en salud pública: Retos antes de la próxima emergencia/pandemia

Pese a la evidente importancia que tiene el contar con un lago de datos con información oportuna y certera para la próxima pandemia, su implementación aún enfrenta múltiples retos que van desde lo estructural hasta lo humano:

1. Barreras estructurales

Los retrasos estructurales en la disponibilidad de las fuentes desde el momento mismo en que se generan los datos son un factor crucial que dificulta la implementación de un lago como este. Aunque se han realizado grandes esfuerzos en las entidades que crean y proporcionan los datos necesarios, cada una opera bajo esquemas ligeramente distintos, con niveles de madurez tecnológica dispares y procesos de verificación costosos, lo que dificulta una integración oportuna y homogénea en tiempo real.

2. El papel clave de la gobernanza

La gobernanza, entendida como la definición clara de los roles, responsabilidades y mecanismos de coordinación entre los distintos actores involucrados no está definida bajo algún marco normativo claro. Sin esta estructura, no es posible realizar una articulación efectiva y sostenida en el tiempo. Asimismo, no existe aún un esquema de costos que distribuya de manera equitativa los esfuerzos requeridos ni que garantice beneficios compartidos entre las partes potencialmente interesadas (Estado-Academia), lo que representa una barrera para la sostenibilidad y el compromiso a largo plazo.

3. Inversión e interoperabilidad

Implementar una solución de datos en tiempo real para futuras emergencias sanitarias implicaría no solo inversiones significativas en infraestructura y automatización de flujos de datos, sino también cambios profundos en políticas institucionales y acuerdos de interoperabilidad entre partes interesadas.

4. Capacidades humanas sostenibles

El mantenimiento de esta infraestructura requiere inversiones constantes y personal altamente calificado. Se necesita contar con profesionales que no solo dominen herramientas de gestión y análisis de datos, sino que también comprendan las fuentes de información en salud pública y sus contextos. Este personal debe garantizar la calidad, trazabilidad y uso eficiente del lago de datos, respondiendo con agilidad ante nuevas demandas durante una emergencia sanitaria. En la actualidad esta capacidad humana en el país es escasa y está limitada.

Recomendaciones

1. **Fortalecer la gobernanza nacional de datos en salud.** Desarrollar un marco regulatorio que incluya estándares técnicos de interoperabilidad, criterios de anonimización, protocolos de seguridad y trazabilidad, y que garantice el acceso seguro a investigadores y tomadores de decisiones.
2. **Asegurar sostenibilidad a largo plazo.** Invertir en infraestructura tecnológica escalable y recursos humanos capacitados, tanto en instituciones públicas como en la academia, para garantizar la operación y actualización permanente de los lagos de datos.
3. **Impulsar mecanismos de colaboración e intercambio de datos.** Establecer alianzas formales y constantes entre el Ministerio de Salud, entidades territoriales, universidades y centros de investigación, con esquemas que promuevan el uso compartido de datos bajo principios éticos, regulatorios y de seguridad.
4. **Incorporar lagos de datos en la preparación y respuesta ante emergencias.** Integrar plataformas como ÁGORA en los sistemas de vigilancia en salud pública y planes de contingencia, permitiendo análisis en tiempo real, simulaciones de escenarios y evaluación del impacto de intervenciones en todo el territorio nacional.

Autores del informe

Andrés Moreno Barbosa

Andrés Moreno Barbosa es profesor asistente en el Departamento de Ingeniería de Sistemas de la Pontificia Universidad Javeriana. Ingeniero y magíster en Ingeniería de Sistemas y Computación y Doctor en Informática. Cuenta con una sólida trayectoria en inteligencia artificial, sistemas de recomendación y ciencia de datos aplicada. Ha trabajado en proyectos de analítica avanzada para la toma de decisiones y colabora en iniciativas de investigación que integran inteligencia artificial y análisis de grandes volúmenes de datos.

Jennifer Murillo Alvarado

Estadística y magíster en epidemiología de la Universidad del Valle. Desde el Instituto de Salud Pública de la Pontificia Universidad Javeriana, aplica su experiencia en la investigación de enfermedades infecciosas, el análisis de resultados en salud, el desarrollo de modelos matemáticos adaptados a Colombia en su comprensión de la dinámica entre el cambio climático y el dengue, para la generación de evidencia y su contribución en el análisis para la generación de evidencia sobre COVID-19 su Respuesta y Lecciones Aprendidas para la Post-Pandemia y Futuras Epidemias.

Daniel Bonilla

Ingeniero mecatrónico de la Universidad Nacional de Colombia; magíster en ingeniería electrónica de la Pontificia Universidad Javeriana (PUJ). Su trayectoria académica se complementa con una amplia experiencia docente en el Departamento de Electrónica de la PUJ. Su campo de acción abarca el diseño y control de sistemas robóticos, el procesamiento avanzado de imágenes y señales biomédicas, y el desarrollo de algoritmos de optimización y aprendizaje de máquina para aplicaciones en salud y neurociencias. Actualmente, combina su labor docente, investigativa y de liderazgo en proyectos académicos e interdisciplinarios que aplican la ciencia de datos en medicina.

Johan Calderón R.

Biólogo con maestría en ecología y doctorado en ecología de enfermedades, con experiencia investigativa en eco-epidemiología, bioestadística, análisis espaciales, y ciencia de datos biológicos. En el proyecto ÁGORA participó en la modelación basada en agentes de la transmisión de SARS-CoV-2 en los diferentes departamentos de Colombia incluyendo Bogotá, la creación de un lago de datos con información de la atención en salud en todo el país, y la elaboración de diferentes visualizaciones con información cuantitativa relacionada con la pandemia de COVID-19.

Jenny Pinilla

Médica y Epidemióloga con amplia experiencia en Salud Pública, en diferentes campos incluyendo planeación, organización y gestión de servicios de salud, vigilancia epidemiológica y trabajo con comunidades urbanas, rurales e indígenas. Desde los cargos desempeñados he realizado implementación de programas de protección específica y detección temprana, a partir del reconocimiento de características sociales, demográficas y de salud de la población; y de la implementación de estrategias innovadoras para el logro de resultados en la gestión y atención a personas con Condiciones Crónicas y promoción de la Actividad Física en contextos urbanos.

Juan G. Torres

Actualmente es el coordinador del centro de cómputo de alto desempeño HPC-ZINE en la Universidad Javeriana. Es Ingeniero Electrónico de la Universidad del Valle (2007) con una MSc. en Ing. Electrónica y Computación (2014) y Ph.D. en Ingeniería (2020) de la Universidad de los Andes. Cuenta con experiencia en el sector industrial, docencia y grupos de investigación en Ingeniería por más de 14 años en áreas de telecomunicaciones, regulación de espectro, computación paralela y distribuida. Dentro de sus áreas de interés ha profundizado en temáticas como estadística, probabilidad y procesos estocásticos para el modelamiento de propagación de ondas electromagnéticas y planificación del espectro electromagnético.

Autores del informe

Jaime A.
Pavlich-Mariscal

Profesor Titular de la Pontificia Universidad Javeriana (Colombia). Sus áreas de interés son: Ingeniería de Software, Software Científico, Análisis de Datos y Procesamiento del Lenguaje Natural. Ha publicado varios artículos sobre dichos temas en conferencias y revistas. A lo largo de su carrera, ha desempeñado roles como desarrollador, arquitecto y consultor especializado en proyectos de ingeniería de software para diversas organizaciones, incluidas universidades, entidades gubernamentales y empresas del sector privado, tanto a nivel nacional como internacional.

Zulma
Cucunubá

Médica de la Universidad Pedagógica y Tecnológica de Colombia (UPTC); magister en Salud Pública de la Universidad Nacional de Colombia; PhD en Epidemiología de enfermedades infecciosas del Imperial College London donde también realizó su posdoctorado en modelamiento de infecciones globales y vacunas. Ha sido profesora del Departamento de Epidemiología Clínica y Bioestadística y profesora Honoraria en el Centro MRC para el Análisis de Enfermedades Infecciosas Globales en el Imperial College London. Su investigación se centra en aplicar modelos estadísticos y matemáticos para estudiar la propagación de enfermedades infecciosas y evaluar la efectividad de intervenciones, con un interés particular en América Latina. Actualmente es la Directora del Instituto de Salud Pública de la Pontificia Universidad Javeriana.

Proyecto ÁGORA

ÁGORA: “Alianza para la Generación de evidencia sobre Covid-19, su respuesta y lecciones Aprendidas para la postpandemia y futuras epidemias”, financiado por el Ministerio de Ciencia, Tecnología e Innovación de Colombia. El proyecto fue ejecutado por la Pontificia Universidad Javeriana en colaboración con la Universidad de los Andes, la Universidad Industrial de Santander, la Universidad del Rosario, el Instituto de Evaluaciones Tecnológicas en Salud y la Cuenta de Alto Costo.

FINANCIACIÓN EJECUCIÓN COLABORACIÓN