

CIENCIA DE DATOS

SEORL-CCC

Big Data e
Inteligencia Artificial



LA CIENCIA DE DATOS	3
LA CIENCIA DE DATOS EN OTORRINOLARINGOLOGÍA Y CIRUGÍA DE CABEZA Y CUELLO	4
Características de la especialidad	4
Definición del problema	5
Objetivos de la ciencia de datos en Otorrinolaringología y Cirugía de Cabeza y Cuello	5
Objetivo general	5
Objetivos específicos	6
Recopilación de datos	6
Limpieza y preparación de datos	7
Análisis exploratorio de datos	7
Modelado de datos	7
Validación y evaluación de modelos	8
Visualización y comunicación	8
Implementación y monitorización	8
RECURSOS NECESARIOS PARA DESARROLLAR CIENCIA DE DATOS EN OTORRINOLARINGOLOGIA Y CIRUGÍA DE CABEZA Y CUELLO	9
Recursos de estructura	9
Organización	
Profesionales	
Equipamiento y dotación	
Recursos de proceso	10
Protocolo de participación en la ciencia de datos de la SEORL-CCC	
Protocolo de idoneidad en los centros para su participación en la ciencia de datos de la SEORL-CCC	
Protocolo de desarrollo de la ciencia de datos en la SEORL-CCC	
Recursos de resultado	11

La ciencia de datos

La ciencia de datos es un campo interdisciplinar que **convierte datos en conocimiento útil para entender el mundo** a través de los datos y ayudar a las organizaciones a tomar decisiones más informadas y resolver problemas complejos. Los datos están en todas partes, y saber usarlos puede marcar la diferencia en cualquier campo.

La ciencia de datos es el conjunto de **herramientas, técnicas y conocimientos** que permiten acceder a este conocimiento. Utiliza métodos, procesos, algoritmos y sistemas científicos para extraer conocimiento e *insights* (información valiosa) a partir de datos estructurados y no estructurados. Es un campo interdisciplinario que combina habilidades y conocimientos de distintas áreas:

- La **estadística** permite analizar y entender los datos.
- La **informática** permite manejar grandes volúmenes de datos y programar algoritmos.
- Las **matemáticas** sirven para crear modelos que expliquen o predigan comportamientos.
- **Conocimiento del área de trabajo o dominio**, para entender el contexto y aplicar los resultados de manera útil.

El proceso de la ciencia de datos suele seguir estos pasos:

1. Definir el problema
2. Recopilación de datos: obtener la información necesaria a partir de bases de datos, equipos tecnológicos, cuestionarios, encuestas, etc.
3. Limpieza y preparación de datos: los datos suelen estar desordenados, incompletos o con errores. Aquí se organizan y se procesan para eliminar errores, inconsistencias o valores faltantes, y transformarlos en un formato adecuado para su análisis.
4. Análisis exploratorio de datos: explorar los datos para entender qué hay en los datos, identificar patrones, tendencias, relaciones y anomalías mediante técnicas visuales y estadísticas.
5. Modelado de datos: aplicar algoritmos de aprendizaje automático (*machine learning*) y técnicas estadísticas para construir modelos predictivos o descriptivos.
6. Validación y evaluación: probar la precisión y eficacia de los modelos utilizando métricas y técnicas de validación.
7. Visualización y comunicación: presentar los resultados de manera clara y comprensible mediante gráficos, paneles interactivos o informes, para que los responsables y equipos implicados puedan tomar decisiones informadas.
8. Implementación y monitorización: desplegar los modelos en entornos productivos y supervisar su rendimiento para asegurar que sigan siendo útiles y precisos con el tiempo.

La ciencia de datos permite:

- La toma de decisiones: ayuda a las organizaciones a tomar decisiones basadas en datos, no en suposiciones.

- Automatización: permite crear sistemas inteligentes que hacen cosas como recomendar pruebas diagnósticas, tratamientos, distribución de recursos.
- Innovación: descubre patrones y oportunidades que antes no eran visibles.
- Eficiencia: optimiza procesos, reduce costes y mejora la experiencia del usuario.

Hay un amplio campo de aplicación de la ciencia de datos en medicina:

- Predicción de tendencias en la evolución de enfermedades.
- Recomendaciones personalizadas en función de biomarcadores.
- Diagnóstico médico y análisis de imágenes en salud.
- Detección de errores e ineficiencias.
- Optimización de protocolos y procesos asistenciales.
- Análisis de comportamientos de pacientes en cuestiones como la adherencia a los tratamientos o la utilización de recursos.

Se utilizan una serie de herramientas y tecnologías comunes:

- Lenguajes de programación: Python, R, SQL.
- Herramientas de visualización: Tableau, Power BI, Matplotlib, Seaborn.
- Marcos de aprendizaje automático (machine learning): Scikit-learn, TensorFlow, PyTorch.
- Plataformas de big data: Hadoop, Spark.
- Bases de datos: MySQL, PostgreSQL, MongoDB.
- Lenguajes de datos y estadística: R.

La ciencia de datos en Otorrinolaringología y Cirugía de Cabeza y Cuello

Características de la especialidad

La especialidad de Otorrinolaringología y Cirugía de Cabeza y Cuello se caracteriza principalmente por 3 condiciones:

1. La **imagen**: es una especialidad eminentemente visual que utiliza la endoscopia y la microscopia como herramientas básicas de exploración, diagnóstico y seguimiento de los pacientes. Igualmente se apoya en las imágenes radiológicas de todo tipo para el diagnóstico, la planificación preoperatoria y el seguimiento de los pacientes. Por su frecuencia, las fuentes de imagen radiológica más utilizadas en la especialidad son la Tomografía Computarizada y la Resonancia Magnética Nuclear.
2. **Los datos cuantitativos y semicuantitativos**: la especialidad utiliza herramientas exploratorias para el diagnóstico y el seguimiento de muchos pacientes recogiendo datos cuantitativos aportados por dispositivos tecnológicos de alta complejidad propios de la

especialidad como: los datos audiológicos de todo tipo con unidades de medida en dB, Hz y ms; los datos de la voz medidos en Hz, dB y s; los datos de escalas de todo tipo (NPS, LOEM, NOSE, BOT8, EVAs...).

3. **Los datos cualitativos:** la especialidad de otorrinolaringología ha asumido plenamente la importancia de las opiniones de los pacientes para evaluar sus estados de salud y para adoptar decisiones sobre pruebas diagnósticas y tratamientos. Se utilizan numerosos cuestionarios y escalas que cuantifican numéricamente estas evaluaciones.

Definición del problema

La especialidad de Otorrinolaringología **genera grandes y valiosas cantidades de datos clínicos, imagenológicos, cuantitativos y cualitativos que no se están recopilando ni explotando** adecuadamente, por lo que requieren una sistematización para su aprovechamiento clínico y científico.

Se está incrementando el desarrollo de proyectos de investigación a pequeña escala sobre aspectos muy concretos de la especialidad que involucran a un solo centro o a un reducido número de investigadores o de centros participantes. Pero se trata realmente de **small data**. La ciencia de datos en estos casos está escasamente profesionalizada por parte de otorrinolaringología, con una gran carga del trabajo depositada en los propios otorrinolaringólogos investigadores o en los residentes. La pervivencia en el tiempo de este modelo local es reducida, vinculada a la voluntad de los investigadores, a su capacidad de dedicar tiempo a las tareas y al cambiante apoyo de los residentes.

La aspiración de la SEORL-CCC es la de crear escenarios operativos de **big data** que perduren en el tiempo.

Objetivos de la ciencia de datos en Otorrinolaringología y Cirugía de Cabeza y Cuello

Objetivo general

Diseñar e impulsar un modelo profesionalizado y sostenible de ciencia de datos, apoyado en equipos técnicos, infraestructura digital segura y coordinación desde una unidad central.

Objetivos específicos

1. Inculcar entre los otorrinolaringólogos la idea de **considerar a la ciencia de datos como un campo verdaderamente interdisciplinar** en el que el especialista en otorrinolaringología aporta su enorme conocimiento sobre el área de trabajo donde se deben aplicar sus preguntas de investigación, pero donde participan también profesionales de otras disciplinas como las matemáticas, la informática y la estadística. La misión del otorrinolaringólogo es la de identificar las necesidades de investigación y dirigir los proyectos de investigación.

2. Transmitir a los otorrinolaringólogos el mensaje de que **hoy día no es posible la autosuficiencia del investigador otorrinolaringológico aislado** para ser competente todos los campos de la ciencia que involucran a una investigación.
3. Convencer a los otorrinolaringólogos de que ni los otorrinolaringólogos investigadores ni los residentes son la fuerza de trabajo que recopila los datos y los introduce en las bases de datos. **Las tareas de bajo valor profesional deben ser llevadas a cabo por personal de apoyo** profesionalizado para esas tareas.
4. Proponer un **modelo de trabajo en ciencia de datos de tipo colaborativo multicéntrico** con el que se pueda alcanzar un auténtico escenario de **big data**. La Sociedad Española de Otorrinolaringología y Cirugía de Cabeza y Cuello se presenta como la entidad idónea para dar soporte a este modelo.
5. Extender el modelo de trabajo en ciencia de datos para que pueda ser utilizado en el nivel local cuando solo se requiere la utilización de **small data**.
6. Crear la **infraestructura necesaria en la SEORL-CCC** para albergar la gestión de los datos mediante servidores informáticos y demás equipamiento básico que permitan la recopilación de datos y toda su posterior gestión.
7. Disponer en la SEORL-CCC de **profesionales multidisciplinares de la ciencia de datos** que lleven a cabo las funciones y tareas de la gestión de los datos que puedan requerirse en todas sus fases.
8. Elaborar desde la SEORL-CCC un **modelo de financiación de proyectos de investigación** en otorrinolaringología y cirugía de cabeza y cuello que utilicen la ciencia de datos basados en el rigor, la transparencia, la robustez metodológica y la obtención de resultados.
9. **Utilizar el modelo de acreditación de Unidades de la SEORL-CCC** en las distintas áreas de la especialidad como vehículo para desarrollar el modelo de ciencia de datos.
10. Utilizar la ciencia de datos como una herramienta de **empoderamiento** de la especialidad de Otorrinolaringología y Cirugía de Cabeza y Cuello y facilitar su **posicionamiento** como una especialidad de vanguardia en el campo de la ciencia.

Recopilación de datos

La recogida y recopilación de datos es el paso fundamental en la cadena de la ciencia de datos. Los otorrinolaringólogos tenemos una profunda laguna de conocimiento en la elaboración de tablas de recogida de datos y en elaboración y utilización de bases de datos. La inmensa mayoría de las tablas que se utilizan en los distintos centros no están bien construidas, no tienen formato de tabla, carecen de celdas con textos desplegable para estandarizar los contenidos, no contemplan la solución de los datos faltantes (*missing values*) y no están orientadas hacia la explotación con herramientas de inteligencia artificial.

El modelo de ciencia de datos de la SEORL-CCC es consciente de estas deficiencias y elabora un **programa formativo en recogida de datos** para que los distintos Servicios de ORL dispongan de tablas y bases de datos adecuadas a la ciencia de datos.

El modelo de ciencia de datos de la SEORL-CCC propone la **financiación de proyectos de investigación** en otorrinolaringología y cirugía de cabeza y cuello que incluyan tanto la elaboración de tablas y bases de datos apropiadas como la **contratación de personal** para que efectúe operativamente la recogida de datos.

Limpieza y preparación de datos

A pesar de hacer una recogida de datos completa y exhaustiva, las tablas y las bases de datos contienen errores, datos desordenados, inconsistencias o valores faltantes.

El modelo de ciencia de datos de la SEORL-CCC plantea una limpieza y preparación de datos profesionalizada, organizando la información recogida y procediendo a su **procesado para transformar los datos en un formato adecuado para su análisis**.

Análisis exploratorio de datos

Solo desde un profundo conocimiento de las herramientas de la ciencia de datos es posible explorar los datos para entender qué hay en los datos, identificar patrones, tendencias, relaciones y anomalías mediante técnicas visuales y estadísticas.

El modelo de ciencia de datos de la SEORL-CCC plantea que sean **profesionales expertos quienes realicen los análisis exploratorios de los datos**.

Modelado de datos

La aplicación de **algoritmos de aprendizaje automático (*machine learning*)** y técnicas estadísticas para construir modelos predictivos o descriptivos a partir de los datos supera las capacidades habituales de los otorrinolaringólogos. Sería un error que los otorrinolaringólogos tuvieran que aprender nuevos lenguajes de programación y nuevos métodos estadísticos. Se trata de campos sometidos a una evolución extraordinariamente rápida que su dominio está restringido a profesionales que solo se dedican a ellos.

No es posible pretender que los otorrinolaringólogos aprendamos los nuevos conocimientos y actualizaciones de nuestra propia especialidad y, además, todos los conocimientos y herramientas de la ciencia de datos. No disponemos del tiempo para ello ni tampoco debemos asumirlo como una obligación sin la que es imposible investigar.

El modelo de ciencia de datos de la SEORL-CCC plantea que sean **profesionales expertos quienes realicen el modelado de los datos**.



Validación y evaluación de modelos

Los modelos predictivos o descriptivos que se pueden haber construido durante la investigación necesitan validarse y se debe probar su precisión y eficacia. Las métricas y técnicas de validación superan igualmente las capacidades de los otorrinolaringólogos para llevar a cabo con éxito este paso de la ciencia de datos. Sí es cierto que algunas validaciones y evaluaciones de los modelos van a requerir nuevos trabajos de campo donde los otorrinolaringólogos son partícipes de es-te paso durante su actividad asistencial.

El modelo de ciencia de datos de la SEORL-CCC plantea que sean **profesionales expertos quienes realicen la validación y evaluación de los modelos.**

Visualización y comunicación

El paso de presentar los resultados de manera clara y comprensible mediante **gráficos, paneles interactivos o informes** es algo propio a los otorrinolaringólogos, lo hemos hecho siempre. La gran diferencia es el camino que se ha seguido para llegar a este punto.

El camino habitual ha sido partir de un trabajo local de investigado-res vocacionales que han quitado horas, días y meses de su vida para intentar llevar a cabo todas las fases de una investigación. Las carencias competenciales en la gestión del dato, según se requiere actualmente por la ciencia de datos, hacen que sus resultados tengan un alcance limitado y sometido a incertidumbres por su reducido número de casos. Algunas de las soluciones han sido extender la participación a otros centros para aumentar esta "n", pero incluso las investigaciones con números "n" aparentemente altos no pueden competir con las investigaciones que incluyen cientos de miles de casos.

El modelo de ciencia de datos de la SEORL-CCC plantea que la visualización y comunicación de los resultados de las investigaciones provenga de la ejecución de un proceso plenamente integrado en la ciencia de datos. El avance del conocimiento se sostendrá con mucha más solidez en estas visualizaciones y comunicaciones y servirá para que los responsables y equipos implicados puedan tomar decisiones informadas.

Implementación y monitorización

Este último paso de la ciencia de datos cierra el círculo del proceso al necesitar que **los otorrinolaringólogos desplieguen los modelos en sus entornos clínicos y supervisen su rendimiento** para asegurar que sigan siendo útiles y precisos con el tiempo.

Recursos necesarios

Recursos de estructura

Organización:

- Se requiere disponer de un modelo de ciencia de datos dirigido y sostenido por la SEORL-CCC.
- Se requiere disponer de un modelo dual en el que la SEORL-CCC se sitúe como centro organizador de *big data* y que, además, preste apoyo a los centros periféricos que desarrollen proyectos de *small data*.
- Se requiere extender el modelo de acreditación de Unidades de la SEORL-CCC.

Profesionales:

- Se requiere disponer de un profesional local en cada centro que actúe como *data manager* o *data entry* vinculado a la estructura de investigación oficial del centro y que preste servicios de recopilación de datos a la Unidad acreditada o en proceso de acreditación del Servicio de ORL.
- Se requiere disponer de profesionales administrativos en la Secretaría de la SEORL-CCC para gestionar la documentación contractual, informativa y de contenidos y toda la vinculada a la gestión de agendas de profesionales locales y remotos, además de la gestión de envíos.
- Se requiere disponer de profesionales multidisciplinares con competencias relacionadas con la ciencia de datos: matemático, informático, estadístico.

Equipamiento y dotación:

- Se requiere disponer en el nivel local de tablas para la recopilación de datos con formato de tabla, apropiadas para la explotación de la información mediante herramientas de inteligencia artificial.
- Se requiere disponer de herramientas de software, estadísticas y de programación adecuadas y con licencia para el desarrollo de la ciencia de datos de la SEORL-CCC. Concretamente, se requiere disponer de lenguajes de programación (Python, R, SQL), herramientas de visualización (Tableau, Power BI, Matplotlib, Seaborn), frameworks de machine learning (Scikit-learn, TensorFlow, PyTorch), plataformas de *big data* (Hadoop, Spark), bases de datos (MySQL, PostgreSQL, MongoDB).
- Se requiere disponer de un sistema de transferencia y compartición de la información entre centros y con el centro de datos de la SEORL-CCC con una metodología de flujo de trabajo para diseñar bases de datos de investigación y con capacidad para ejecutar procedimientos automatizados de exportación para descargas de datos sin problemas a Excel, PDF y paquetes estadísticos comunes (SPSS, SAS, Stata, R).

Se requiere que este sistema disponga de rutas de auditoría para rastrear la manipulación de datos y la actividad del usuario.

- Se requiere disponer de servidores informáticos capaces de almacenar toda la información en modo seguro y que cumpla todos los protocolos de seguridad frente a riesgos informáticos y de protección de datos de carácter personal.

Recursos de proceso

Protocolo de participación en la ciencia de datos de la SEORL-CCC:

- Se requiere que la Unidad participante en el modelo de ciencia de datos se encuentre acreditada por la SEORL-CCC o en proceso activo de acreditación.
- Se requiere establecer una relación contractual entre la SEORL-CCC y las estructuras oficiales de investigación de cada centro.
- Se requiere que la Unidad del Servicio de ORL participante en el modelo de ciencia de datos de la SEORL-CCC disponga de un proyecto de investigación adecuadamente estructurado y aprobado por el Comité de Ética de la Investigación de su centro.

Protocolo de idoneidad de los centros para su participación en la ciencia de datos de la SEORL-CCC:

- Se requiere disponer de un Acuerdo de compartición y transferencia de datos entre la SEORL-CCC y la estructura de investigación local.
- Se requiere disponer de un contrato de encargado de tratamiento de datos de carácter personal en el nivel local.
- Se requiere disponer de un protocolo de recogida y recopilación de datos que cumpla las recomendaciones de *Datasheets for Datasets*¹.
- Se requiere disponer de un protocolo que cumpla los principios rectores de FAIR para la gestión y administración de datos científicos ².

¹ Gebru Timnit, Morgenstern Jamie, Vecchione Briana, Vaughan Jennifer Wortman, Wallach Hanna, Daumé III Hal, Crawford Kate. Datasheets for Datasets. Disponible en: <https://arxiv.org/pdf/1803.09010>

² Gebru Timnit, Morgenstern Jamie, Vecchione Briana, Vaughan Jennifer Wortman, Wallach Hanna, Daumé III Hal, Crawford Kate. Datasheets for Datasets. Disponible en: <https://arxiv.org/pdf/1803.09010>



Protocolo de desarrollo de la ciencia de datos en la SEORL-CCC:

- Se requiere disponer de un protocolo detallado de las actividades y tareas que deben llevarse a cabo en cada paso de la ciencia de datos: limpieza y preparación de datos; análisis exploratorio de datos; modelado de datos; validación y evaluación de modelos; visualización y comunicación de los resultados; implementación y monitorización de los modelos.
- Se requiere que este protocolo dé respuestas a las siguientes preguntas en cada paso de la ciencia de datos: ¿qué?, ¿quién?, ¿cuándo?, ¿dónde?, ¿cómo?, ¿con qué? y ¿por qué?

Recursos de resultado

- Se requiere disponer de herramientas de inteligencia artificial para la visualización y difusión de los resultados.
- Se requiere disponer de financiación para la difusión y divulgación de los resultados.
- Se requiere disponer de un protocolo destinado a ordenar la participación de los investigadores locales de cada centro y establecer los acuerdos de difusión y divulgación de los resultados en cuanto a la distribución de las autorías y asistencia a eventos científicos para su presentación.