**1) En el campo de la investigación médica, ¿cómo y para qué se utiliza la programación? ¿Hasta qué punto crees que es necesaria?**

La investigación biomédica tiene múltiples disciplinas que abarcan desde investigación fundamental (genómica, proteómica, mecanismos moleculares, etc.) hasta los registros y ensayos de investigación clínica con pacientes. En todos estos campos, la programación es una herramienta básica. El análisis estadístico de los resultados es un aspecto común a estas disciplinas, que cada vez está adquiriendo mayor complejidad. Y la mayor garantía de trazabilidad y reproducción de resultados se hace mediante programación estructurada de los análisis estadísticos. En campos de la investigación básica como la genómica, metagenómica, transcriptómica o proteómica, la programación es imprescindible para el tratamiento de la gran cantidad de información que se genera. En los estudios observacionales o experimentales en el ámbito clínico, el procesado de datos de historia clínica electrónica, la depuración y minería de datos también exige de conocimientos avanzados de programación. Por este motivo, la investigación biomédica moderna exige, en la inmensa mayoría de los casos, de participación en los equipos de personal con conocimientos de programación.

**2) En tu caso, ¿por qué escogiste R como herramienta de trabajo y no otros lenguajes de programación?**

Yo personalmente llegué a R desde su versión primordial (comercial) S-Plus, que era el lenguaje de análisis estadístico que utilizaba a principios de los 2000. R tiene enormes ventajas como que es software libre y gratuito, tiene muchísimas librerías de paquetes con los análisis más sofisticados, incorpora capacidades gráficas muy avanzadas y en los últimos años ha avanzado mucho integrando librerías muy avanzadas de inteligencia artificial. Mediante R es posible llevar a cabo prácticamente todos los aspectos a los que me refería arriba, desde la lectura de los datos hasta la generación de tablas, figuras e incluso informes completos (con Markdown). Y mediante la utilización y consolidación del código de análisis, todo queda perfectamente trazable desde el principio hasta el final. Este aspecto es fundamental en la investigación biomédica, y los programas de análisis estadísticos basados en interfaz gráfica son más limitados en este sentido (como, por ejemplo, STATA).

Es cierto que tiene una curva de aprendizaje bastante pronunciada al principio, y es difícil para profesionales que tienen menor formación en programación estructurada. Pero una vez que se comprende su estructura básica y la programación por objetos, el potencial es enorme. Posiblemente, en estos ámbitos, R y Python son las dos herramientas de más desarrollo en el futuro.

**3) ¿Podrías contarnos más acerca del uso de R en alguno de tus proyectos?**

En el equipo en el que trabajo utilizamos R de forma rutinaria. Ahora mismo estamos trabajando en un estudio prospectivo basado en explotación de datos de historia clínica electrónica que pivota todo alrededor de distintas rutinas en R. Primero hacemos procesado de lenguaje natural (NLP- natural language processing) para extracción y codificación de la información contenida en informes clínicos. También utilizamos herramientas de SQL incorporadas en R para acceder a la información de historia clínica electrónica. Después utilizamos R para toda la depuración de datos y análisis posterior de supervivencia. Estamos utilizando técnicas estadísticas convencionales (regresión múltiple, riesgos proporcionales, etcétera), pero también aproximaciones de Deep-learning. Para ello hemos montado un servidor dedicado con hardware específico en el que entrenaremos los modelos de predicción. El objetivo final es, mediante la explotación de información clínica y de imagen cardiaca, poder mejorar la predicción de riesgo asociado a una enfermedad conocida como “estenosis aórtica”, que es muy prevalente y grave, pero que tratada a tiempo tiene excelente pronóstico. El objetivo es explotar al máximo la información clínica ya disponible en los hospitales para intentar afinar con más precisión el momento óptimo de tratamiento. En la Figura muestro cómo el riesgo de morirse (representado como el “Hazard Ratio” que es cuál es la probabilidad de morirse comparada con un paciente control con los valores de referencia) varía en función de dos índices que se obtienen mediante imagen cardiaca (ecocardiografía): la velocidad máxima de la sangre a través de la válvula (Vmax) y el tamaño del orificio efectivo de la válvula por la que sale la sangre del ventrículo izquierdo (AVA):



Estos análisis están basados en más de 20.000 ecocardiogramas y la información de mortalidad que extraemos de los registros de la Comunidad de Madrid. Todo el proceso está hecho análisis en R: desde el NLP hasta las gráficas que se muestran (utilizando la librería ggplot).

**4) Por último, ¿cómo crees que se debería enseñar a programar (en R y en otros lenguajes)?**

Realmente en esto es posible que tenga una visión un tanto particular, debido al hecho de que he sido fundamentalmente autodidacta. Creo que una formación estructurada es útil para coger buenos hábitos desde el principio. La algoritmia y empezar a pensar la estructura de los programas antes de empezar a escribir código es altamente recomendable, y en ello debe invertirse un buen esfuerzo desde el principio. Con esta buena base, después las características de sintaxis y semántica específicas de cada lenguaje no son tan difíciles de cambiar.

Mi consejo es aspectos fundamentales de la formación técnica en investigación biomédica (análisis estadístico, procesado de imagen médica o bioinformática de datos moleculares) se aprenda de la mano la parte teórica con casos prácticos concretos en los que se trabaje con datos reales. Mi consejo es que se acople la formación en programación con la realización de trabajos prácticos que se hagan de forma paralela a la adquisición de conocimientos teóricos.