APLICACION DE TÉCNICAS DE MACHINE LEARNING PARA LA PREDICCIÓN DE RIESGO DE ALERGIA A ANTIBIOTICOS BETALACTÁMICOS

ALICIA GALLARDO HIGUERAS

DIRECTORES:

VIDAL MORENO RODILLA
ESTHER MORENO RODILLA

PLAN DE INVESTIGACIÓN

PROGRAMA DE DOCTORADO FORMACIÓN EN LA SOCIEDAD DEL CONOCIMIENTO

UNIVERSIDAD DE SALAMANCA

FECHA: 27 DE MAYO DE 2022

INTRODUCCIÓN

La Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) establece un sistema de farmacovigilancia de medicamentos de uso humano. Es muy importante reconocer y diagnosticar las reacciones adversas medicamentosas graves que puedan amenazar la vida del paciente, provoquen su hospitalización o la prolonguen, ocasionen incapacidad laboral o escolar, induzcan defectos congénitos o sean clínicamente relevantes.

Según la AEMPS, una reacción adversa (RAM) "es cualquier respuesta nociva y no intencionada a un medicamento" y también "es cualquier suceso indeseable que ha sucedido con el paciente mientras estaba utilizando un medicamento y existe la sospecha de que es causado por el medicamento".

Las reacciones alérgicas inducidas por fármacos se caracterizan por estar mediadas por mecanismos inmunológicos. Esto implica que son específicas del fármaco inductor, remiten al suspenderlo, son reproducibles con la reintroducción del mismo, requieren una exposición previa (sensibilización) a ese fármaco y puede existir reactividad cruzada con otras sustancias de estructura química similar.

En España, casi el 15% de los pacientes que acude a una consulta de Alergia lo hace por una sospecha de hipersensibilidad a un fármaco, siendo el tercer motivo de consulta tras la rinoconjuntivitis y el asma bronquial. Estas reacciones suponen una carga considerable, tanto al paciente como al hospital, dado que, en muchos casos, requieren la realización de varias visitas para concluir el estudio.

Los antibióticos betalactámicos (BL) siguen siendo la primera línea de tratamiento para muchas infecciones bacterianas. Los BL, y sobre todo las penicilinas, son también los fármacos implicados con mayor frecuencia en las reacciones alérgicas por fármacos. Sin embargo, es habitual el sobrediagnóstico, ya que entre un 70-90% de pacientes ambulatorios y hospitalizados etiquetados de alergia a BL realmente no lo son cuando se realiza el estudio alergológico (Romano 2020, Moreno 2016). Las consecuencias de una etiqueta de alergia a la penicilina son significativas, tanto para los pacientes como para los sistemas de salud pública. Numerosos estudios han establecido la carga y el impacto de la alergia a los antibióticos BL. Los pacientes etiquetados de alergia a la penicilina son tratados con frecuencia con antibióticos de amplio espectro, que se asocian con un mayor riesgo de infecciones por microorganismos resistentes (Macy, 2014). Además, este etiquetado de alergia a la penicilina también se ha asociado con mayores costes, fallos en el tratamiento y aumento en los días de hospitalización. (Charnesky, 2011; Solensky, 2014].

El abordaje proactivo del diagnóstico de alergia a los BL tiene un gran impacto en el sistema sanitario, y debe ser una estrategia prioritaria de los programas de optimización de uso de antimicrobianos.

Un diagnóstico rápido y preciso resulta crucial para mejorar el uso correcto de la terapia antibiótica, aumenta la seguridad del paciente y reduce los costes para los sistemas de salud. Un diagnóstico alergológico preciso se basa principalmente en las pruebas cutáneas y las pruebas de exposición controlada con fármacos, herramientas que, en la gran mayoría de los casos, consumen una cantidad apreciable de tiempo y no están exentas de riesgo [Romano 2020]. Además, es posible que las pruebas de alergia no estén disponibles en casos de necesidad urgente de terapia con antibióticos.

En los últimos años se ha asistido a un interés creciente por el desarrollo de clasificaciones de estratificación de riesgo con el objetivo de identificar a los pacientes de bajo riesgo etiquetados como alérgicos a la penicilina, en los que se podrían utilizar otros BL de forma segura. Al respecto, se han diseñado varias guías informatizadas basadas en la historia clínica, [Krisnha, 2017]. La implementación de algunas de estas guías ha permitido incrementar el uso de BL (Blumenthal, 2017).

Igualmente, el mercado electrónico pone a nuestra disposición una amplia gama de sensores y herramientas de detección que pueden ser utilizadas para la adquisición de datos de máquina. Es necesario conocer y caracterizar que dispositivos se ajustan mejor a las distintas aplicaciones y características mecánicas de las herramientas, con el fin de maximizar la eficiencia de la tecnología instalada, con el fin de obtener la información más valiosa posible. Por otra parte, se debe considerar que una fuente de información de gran valor está disponible a través de la base documental de las historias clinicas del servicio.

Una vez se han estudiado y escogido los mejores procedimientos para la adquisición de datos de máquina, el siguiente paso es realizar una caracterización y tratamiento de estos datos, y su procesamiento mediante técnicas de "machine learning" donde la colaboración con el GIR "Robótica y Sociedad" es crucial para el desarrollo del trabajo, para lo cual, es necesario estudiar distintas formas de procesamiento de datos, los antecedentes de las aplicaciones en el ámbito de la medicina, estudiando los distintos resultados y comparándolos con las distintas historias tomadas en diversas condiciones.

Para la realización de este trabajo de procesado de datos, conviene estudiar distintas herramientas software de tratamiento de datos, con el fin de desarrollar una metodología propia que nos ofrezca unos resultados fiables e interpretables.

HIPÓTESIS DE TRABAJO Y PRINCIPALES OBJETIVOS A ALCANZAR

El presente proyecto tiene por objeto desarrollar un sistema de análisis y procesado de datos generados en el Servicio de Alergología del Complejo Hospitalario de Salamanca mediante la realización de consultas de atención a pacientes que son susceptibles de ser alérgicos a los antibióticos betalactámicos.

Para la consecución del objetivo principal, se plantean otros objetivos particulares que se desarrollan de manera secuencial, empezando el conocimiento del estado del arte del machine learning en Medicina, el procesamiento de información textual, los mecanismos de predicción inteligentes basados en las codificaciones de palabras, estudiando las soluciones y métodos utilizados para tal fin.

El objetivo de la propuesta es que la prueba que se realice sea sistemática y que tenga como información de entrada elementos inmediatos y con un elevado nivel de disponibilidad. En este sentido, la historia clínica del paciente es un elemento al que cualquier facultativo en cualquier actividad asistencial tiene acceso. La utilización de medios informáticos en la realización de las historias clínicas hace de su disponibilidad algo inmediato. Se trataría de desarrollar una herramienta, en el ámbito del Machine Learning que, tomando como información la citada historia clínica realice una predicción de alergia a un medicamento (Problema de clasificación) o de al menos de la estimación del riesgo de padecerla (Problema de predicción). Es evidente que el problema surge a la hora de analizar la naturaleza de este tipo de información de entrada. Las historias clínicas son realizadas utilizando el lenguaje propio de los diferentes especialistas médicos (primera dificultad) con diferentes formaciones profesionales, diferente experiencia relacionada con las poblaciones asistenciales, etc. El hecho conciliador es que existen elementos comunes en la elaboración de las historias, el tipo de situaciones enfrentadas, así como los conocimientos médicos que se puedan considerar básicos. Esto hace que se pueda considerar que existe un CORPUS común y suficientemente amplio que analizar y procesar para extraer la información disponible. Existen técnicas de procesamiento (word2vect, CBOW, etc.), desarrolladas en la última década que permiten extraer información significativa de textos libres de un dominio de conocimiento (como es el caso del diagnóstico médico). Una vez realizada esta etapa analítica extractiva es posible desarrollar propuestas de realización de técnicas diagnósticas basadas en Redes Neuronales Clásicas o, en los últimos tiempos, mediante la utilización de Redes Convolucionales, que permiten desarrollar predicciones o estimaciones. Cabe señalar que con las posibilidades de los sistemas computacionales actuales se podría realizar un análisis exhaustivo exploratorio de todas las historias clínicas de los pacientes que acuden a una instalación hospitalaria. Si la herramienta desarrollada tiene un comportamiento predictivo mínimamente correcto se eliminaría el sesgo que se ha comentado previamente para seleccionar al paciente susceptible de padecer una situación de riesgo, como es el caso de la alergia a los antibióticos BL, que tienen un uso generalizado en nuestros hospitales.

Como paso previo a la consecución del objetivo principal, resulta necesario focalizar los esfuerzos en la consecución de los objetivos secundarios consistentes en la construcción de modelos de tratamiento y procesado de datos que se encuentran almacenados en las historias clínicas. Para ellos será necesario:

1.- Evaluar experimentalmente distintos orígenes de datos, con la obtención de las herramientas de procesamiento de los mismos.

2.- Desarrollar e implementar procedimientos de codificación de la información textual, validando su capacidad de representación e incluyendo las herramientas de análisis actuales.

3.- Utilizar los procedimientos de aprendizaje automático para desarrollar prototipos que reproduzcan la actividad diagnóstica del servicio de Alergología.

METODOLOGÍA

El equipo de trabajo estará formado por los miembros del GIR “Robótica y Sociedad” de la Universidad de Salamanca con profesores del Departamento de Informática y Automática con sexenio activo y con una larga experiencia en el campo de la Inteligencia Artificial y sus aplicaciones en Medicina, junto con miembros del Servicio de Alergología del Complejo Hospitalario de Salamanca, así como la estudiante del presente doctorado Alicia Gallardo Higueras.

De acuerdo con el hecho de que se trata de un proyecto completamente integrado en el ámbito en la Inteligencia Artificial moderna y de la aplicación de nuevos procedimientos de diagnóstico médico se plantea la utilización de metodologías ágiles en la definición y ejecución de cada uno los subproyectos en los que se puede dividir cada una de las fases y que se concretan en las tareas.

Cada tarea lleva asociada unas iteraciones (entre 3 y 9) con una duración aproximada de 1 mes. Existe en cada caso un responsable de iteración de cada campo involucrado (Medicina e Inteligencia Artificial).

Además de las reuniones diarias de iteración, dentro de la metodología se plantea la realización de reuniones de revisión y retrospectiva de trabajo con periodicidad mensual. Esta propuesta se lleva desarrollando desde enero de 2021, en el momento en que el Servicio de Alergología pone a disposición de la propuesta un conjunto de Historias Clínicas.

Cabe señalar que se ha de suscribir un convenio según el cual se garantiza todo lo referente a la Protección de los datos individuales de las personas que aparecen. De hecho, el nombre es omitido del análisis del CORE documental por cuanto, además, no aporta absolutamente ninguna información útil para el desarrollo de la propuesta investigadora.

Para el desarrollo de las fases iniciales del proyecto se considera el uso de Matlab, que es una conocida herramienta en la comunidad científica general y de la Inteligencia Artificial en particular. Dispone, entre muchas ventajas, de una avanzada ToolBox de “Text Processing” que permite unos tiempos de desarrollo muy optimizados.
En fases posteriores se utilizará el propio Matlab, así como otras herramientas del ámbito de la materia como son TensorFlow para el aprendizaje de las redes, KNIME para el procesado de los Embeddings, etc.

Este plan de investigación ha sido aprobado por el Comité de Ética de Hospital Universitario de Salamanca.

MEDIOS Y RECURSOS MATERIALES DISPONIBLES

Este trabajo se desarrolla en el programa de Doctorado: Formación en la Sociedad del Conocimiento (García-Peñalvo, 2014), siendo su portal (García-Peñalvo et al., 2019), accesible desde http://knowledgesociety.usal.es, la principal herramienta de comunicación y visibilidad de los avances. En él se irán incorporando todas las publicaciones, estancias y asistencias a congresos durante el transcurso del trabajo

Los medios y recursos materiales que se plantean para este trabajo de tesis Doctoral serían:
- Repositorios de información bibliográfica a los que está suscrita la Universidad de Salamanca, así como aquellos delSACYL
- Ordenadores de procesamiento de información personales para la adaptación y/o maquetación de los datos. Se utilizarían las licencias de software que, en su mayor parte, están licenciadas por la Universidad de Salamanca: MS Office, SPSS, etc
- Sistemas de computación de aprendizaje maquina con licencias de software específico: Matlab con Toolbox de procesamiento de textos y aprendizaje máquina. Estas máquinas están equipadas con equipos de procesamiento paralelo de altas prestaciones NVIDIA y están disponibles en las instalaciones del GIR GROUSAL.

PLANIFICACIÓN TEMPORAL

El proyecto está dividido en 5 Fases, ajustadas a 36 meses:

F1. Estudio estado de la cuestión.
En esta fase se han analizado por una parte los problemas relacionados con el diagnóstico eficaz de las reacciones a antibióticos BL y su efecto en la actividad hospitalaria. Por otra parte, se ha realizado un análisis de las principales soluciones que el Machine Learning ofrece para el uso del Machine Learning en Medicina. En particular, se deben analizar las técnicas de "enmarcado de palabras", más conocido por Word Embedding, que se caracterizan por buscar representaciones matemáticas de las relaciones de las palabras dentro de los documentos de un área buscando en ello cómo las expresiones pueden almacenar conocimiento vertido de una forma más o menos consciente por los profesionales en la documentación clínica. Es evidente que el elemento clave dentro de esta documentación clínica es la Historia Clínica. (Meses 1-6)

F2. Procesado de la Información clínica.
En esta fase se ha realizado el procesado de la información clínica proporcionada por el Servicio de Alergología. Tras analizar los ficheros utilizados se han desarrollado herramientas informáticas para el procesado individualizado de las historias y la extracción de los diferentes campos. Cabe señalar que se trata de información generada por los profesionales del Servicio desde finales del siglo pasado (en formato WordPerfect) por lo que se hace necesario disponer de herramientas para la importación de este formato. Destacar, que el Servicio de Alergología dispone de los informes clínicos en Word, realizados desde 1996 hasta la introducción de la historia electrónica. Esta enorme biblioteca de archivos es una de las grandes ventajas de este proyecto. En la actualidad, los ficheros se almacenan en formato MS Word y Adobe PDF. La potencia de las bibliotecas de Microsoft ha permitido desarrollar utilidades que realizan esta tarea de forma completamente automatizada. Con ello, se consigue un beneficio residual del Servicio pues permite mejorar la disponibilidad de los repositorios de Historias al ser almacenados en un formato actual de MS Word. (Meses 6-15)

F3. Representación y Entreno del sistema de predicción basado en Machine Learning (Text-Processing).
Esta fase es crítica para el logro de unos resultados aceptables de la propuesta investigadora. Se van a considerar las diferentes posibilidades como word2vect (usando CBOW o Skipgram). Cabe señalar que en las pruebas preliminares aparece un conjunto vocabulario superior a 1000 palabras. Asimismo, se presenta una dificultad adicional pues algunos de los modelos tienen una limitación con el idioma (inglés, chino, alemán…). Se realizará una evaluación intrínseca del enmarcado de los conceptos clínicos que aparezcan, con ayuda de los miembros del Servicio de Alergología. Aunque no se puede asegurar, de acuerdo con los resultados, se podrán utilizar sistemas avanzados que utilicen sistemas neuronales recurrentes. Otro aspecto importante que considerar será la visualización de los "word-embeddings" obtenidos, y su posible interpretación en lo que se refiere al conocimiento del dominio por parte de los expertos en Alergología (meses 15-24)

F4. Validación y actualización del proceso de diagnóstico automático (meses 23-32).
El objetivo de esta fase es la validación del modelo de predicción. Se utilizarán los datos provenientes de los nuevos casos presentes en el servicio de alergia y en aquellos de otros hospitales con los que se pueda colaborar. Esta fase es crucial para mostrar la validez del procedimiento, por lo que se plantea como una de las más largas dentro del programa de investigación.

F.5 Documentación y plan de difusión (mes 32 – 36).
Este proyecto de tesis incluye el periodo de redacción de la memoria de tesis, así como de las principales publicaciones que soporten y hagan difusión de sus resultados. Se plantean los trabajos de desarrollo de una herramienta de diagnóstico específica.

REFERENCIAS

I. Banerjee, M. C. Chen, M. P. Lungren and D. L. Rubin, “Radiology report annnotation using Intelligent word embeddings: Applied to multi-institutional chest CT cohort”, Journal of Biomedical Informatics, Vol 77. Pp 11-20. 2017.Doi: doi.org/ 10.1016/j.jbi.2017.11.012.

Blumenthal KG, Shenoy ES, Varughese CA, Hurwitz S, Hooper DC, Banerji A. Impact of a clinical guideline for prescribing antibiotics to inpatients reporting penicillin or cephalosporin allergy. Ann Allergy Asthma Immunol2015; 115:294-300 e2.

Blumenthal KG, Wickner PG, Hurwitz S, Pricco N, Nee AE, Laskowski K, et al. Tackling inpatientpenicillin allergies: Assessing tools for antimicrobial stewardship. J Allergy Clin Immunol 2017; 140:154-61 e6.

Blumenthal KG, Lu N, Zhang Y, Li Y, Walensky RP, Choi HK. Risk of meticillin resistant Staphylococcusaureus and Clostridium difficile in patients with a documented penicillin allergy: population based matched cohort study. BMJ 2018; 361:k2400.

Charneski L, Deshpande G, Smith SW. Impact of an antimicrobial allergy label in the medical record onclinical outcomes in hospitalized patients. Pharmacotherapy 2011; 31:742-7.

García-Peñalvo, F. J. (2014). Formación en la sociedad del conocimiento, un programa de doctorado con una perspectiva interdisciplinar. Education in the Knowledge Society, 15(1), 4-9. https://doi.org/10.14201/eks.11641

García-Peñalvo, F. J. (2022). Developing robust state-of-the-art reports: Systematic Literature Reviews. Education in the Knowledge Society, 23, Article e28600. https://doi.org/10.14201/eks.28600

García-Peñalvo, F. J., Rodríguez-Conde, M. J., Verdugo-Castro, S., & García-Holgado, A. (2019). Portal del Programa de Doctorado Formación en la Sociedad del Conocimiento. Reconocida con el I Premio de Buena Práctica en Calidad en la modalidad de Gestión. In A. Durán Ayago, N. Franco Pardo, & C. Frade Martínez (Eds.), Buenas Prácticas en Calidad de la Universidad de Salamanca: Recopilación de las I Jornadas. REPOSITORIO DE BUENAS PRÁCTICAS (Recibidas desde marzo a septiembre de 2019) (pp. 39-40). Ediciones Universidad de Salamanca. https://doi.org/10.14201/0AQ02843940

F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney and F. Rudzicz, “A survey of word embeddings for clinical text”, Journal of Biomedical Informatics, Vol 100S, 2019. Doi: doi.org/10.1016/j.yjbinx.2019.100057

Krishna MT, Huissoon AP, Li M, Richter A, Pillay DG, Sambanthan D, et al. Enhancing antibioticstewardship by tackling "spurious" penicillin allergy. Clin Exp Allergy 2017; 47:1362-73.

Macy E, Contreras R. Health care use and serious infection prevalence associated with penicillin "allergy" inhospitalized patients: A cohort study. J Allergy Clin Immunol 2014; 133:790-6.

Moreno E, Laffond E, Munoz-Bellido F, Gracia MT, Macias E, Moreno A, et al. Performance in real life of theEuropean Network on Drug Allergy algorithm in immediate reactions to beta-lactam antibiotics. Allergy 2016; 71:1787-90.

Moreno EM, Moreno V, Laffond E, Gracia-Bara MT, Muñoz-Bellido FJ, Macías EM, Curto B, Campanon MV,de Arriba S, Martin C, Davila I. Usefulness of an Artificial Neural Network in the Prediction of β-Lactam Allergy. J Allergy ClinImmunol Pract. 2020;8(9):2974-2982.

Romano A, Atanaskovic-Markovic M, Barbaud A, Bircher AJ, Brockow K, Caubet JC, Celik G, Cernadas J,Chiriac AM, Demoly P, Garvey LH, Mayorga C, Nakonechna A, Whitaker P, Torres MJ. Towards a more precise diagnosis of hypersensitivity to beta-lactams - an EAACI position paper. Allergy. 2020;75(6):1300-1315

Solensky R. Penicillin allergy as a public health measure. Journal of Allergy and Clinical Immunology 2014;133:797-8.