Arch Bronconeumol. 2014;50(3):113–119

ARCHIVOS DE BRONCONEUMOLOGIA

www.archbronconeumol.org

Artículo especial

Minería de textos y medicina: utilidad en las enfermedades respiratorias

David Piedra[a,*], Antoni Ferrer[a,b,c,d] y Joaquim Gea[a,b,c,d]

[a] Instituto de Investigación del Hospital del Mar (IMIM), Barcelona, España
[b] Servicio de Neumología, Hospital del Mar, Barcelona, España
[c] Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, España
[d] CIBERES, ISC III, Bunyola, Mallorca, España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 28 de marzo de 2013
Aceptado el 18 de abril de 2013
On-line el 4 de febrero de 2014

Palabras clave:
Minería de textos
Enfermedades respiratorias
Enfermedad pulmonar obstructiva crónica
Cáncer de pulmón
Paciente respiratorio crítico

R E S U M E N

Cada vez es más habitual disponer de información médica en formato electrónico. Esto incluye tanto artículos científicos como revisiones sobre el manejo clínico e incluso registros de instituciones sanitarias con datos de pacientes. Sin embargo, los instrumentos tradicionales, tanto individuales como institucionales, son poco útiles para seleccionar la información más apropiada en cada caso, sea en el ámbito clínico o en el de la investigación. La llamada «minería» de textos o de datos permite gestionar esa gran cantidad de información, extrayéndola de fuentes diversas mediante sistemas de procesamiento (filtrado y curado), integrándola y permitiendo la generación de nuevo conocimiento. La presente revisión pretende proporcionar una idea general sobre la minería de textos y datos, así como sobre la ayuda que esta técnica bioinformática puede suponer para el ejercicio asistencial de la medicina respiratoria y para la investigación en ese mismo campo.

Text Mining and Medicine: Usefulness in Respiratory Diseases

Keywords:
Text mining
Respiratory diseases
Chronic obstructive pulmonary disease
Lung cancer
Critically ill respiratory patients

A B S T R A C T

It is increasingly common to have medical information in electronic format. This includes scientific articles as well as clinical management reviews, and even records from health institutions with patient data. However, traditional instruments, both individual and institutional, are of little use for selecting the most appropriate information in each case, either in the clinical or research field. So-called text or data «mining» enables this huge amount of information to be managed, extracting it from various sources using processing systems (filtration and curation), integrating it and permitting the generation of new knowledge. This review aims to provide an overview of text and data mining, and of the potential usefulness of this bioinformatic technique in the exercise of care in respiratory medicine and in research in the same field.

Introducción

Las enfermedades respiratorias más prevalentes (asma bronquial, enfermedad pulmonar obstructiva crónica [EPOC], infecciones y cáncer de pulmón) suponen un gran reto para la salud y para los sistemas sanitarios, dado el elevado coste económico y social que representan[1–7]. Por otra parte, con el avance de las tecnologías relacionadas con la información es posible acceder y analizar cantidades ingentes de datos relacionados con la salud y la enfermedad. Así, es común registrar datos de pacientes en formato electrónico, tanto en los hospitales como en la medicina primaria. Esto incluye datos de filiación, pero también los relacionados con los diagnósticos, la gravedad, los resultados analíticos, las pruebas funcionales y la medicación, así como las características de los contactos del enfermo con el sistema sanitario. Disponer de esta gran cantidad de información en formato digital ofrece 3 ventajas importantes: a) mejora su calidad; b) reduce el tiempo marginal de trabajo por parte del personal sanitario, y c) ofrece la posibilidad de utilizar dicha información mediante sistemas automatizados, como la «minería de textos» *(text mining)* o la «minería de datos» *(data mining)*[8,9]. Desde un punto de vista estricto, se diferencian en que la primera obtiene la información a partir de formatos de texto libre, y la segunda, a partir de bases de datos. Un ejemplo pionero de utilización de datos clínicos a nivel institucional fue la plataforma MedLEE[10], un sistema automatizado que permitió la obtención de información relevante a partir de los expedientes clínicos.

* Autor para correspondencia.
Correo electrónico: dpiedra@imim.es (D. Piedra).

0300-2896/$ – see front matter
http://dx.doi.org/10.1016/j.arbres.2013.04.009

Otros ámbitos de enorme importancia que resultan potenciados por los instrumentos informáticos son la formación profesional y la investigación. En estos terrenos existe una enorme cantidad de fuentes de información, difíciles de manejar y seleccionar por el profesional médico. De hecho, existe una gran cantidad de publicaciones tanto en papel como electrónicas que intentan recopilar lo más relevante que se produce en el mundo editorial sobre determinada área del conocimiento médico. Un ejemplo serían las series de UpToDate[11], entre las que se hallan algunas dedicadas a la medicina respiratoria (*Pulmonary, Critical Care and Sleep Medicine* y *Allergy and Immunology*, respectivamente). Algo similar ocurre con los sistemas de alerta, que se reciben por correo electrónico atendiendo a perfiles de usuario predeterminados. También en estos casos existen instrumentos de la minería de textos que permiten ir más allá en la búsqueda, la selección y el procesamiento de información.

Biología y medicina de sistemas

La «biología de sistemas» se ha definido como la ciencia dedicada al estudio sistemático de las interacciones que se producen en los sistemas biológicos. En caso de referirse al caso concreto de las enfermedades humanas se hablaría de «medicina de sistemas», aunque esta última expresión tiene también otras acepciones. Para avanzar en la biología-medicina de sistemas es necesario conectar adecuadamente el conocimiento existente en áreas diversas de la ciencia[12], con el fin de aflorar nuevas propiedades e hipótesis no evidenciables a partir de los enfoques tradicionales. Los instrumentos que facilitan esta labor interdisciplinaria son en gran medida producto del desarrollo de las tecnologías de la información, que permiten el procesamiento de grandes cantidades de datos de origen diverso, desarrollando nuevo conocimiento a partir de ellos. Así, la biología-medicina de sistemas permite por ejemplo establecer modelos matemáticos de enfermedades que integran el conocimiento estructural y fisiopatológico en sus diferentes niveles de complejidad[13]. Entre sus objetivos se halla no solo una mejor comprensión de los procesos nosológicos, sino la posibilidad de simulación de sistemas orgánicos, para generar nuevas aproximaciones diagnósticas y terapéuticas.

La biología-medicina de sistemas utiliza a su vez diversos instrumentos. Por un lado, el objeto principal de esta revisión, la *minería de textos*. Esta puede definirse como un conjunto de técnicas informáticas cuyo objetivo es detectar, extraer e interpretar, de forma automatizada (o semiautomatizada), una información digital que está básicamente en formato de texto. Por otro lado, están los ya mencionados modelos matemáticos de procesos biológicos o nosológicos. Para elaborarlos, deben utilizarse instrumentos y procedimientos capaces de procesar adecuadamente la gran cantidad de información suministrada tanto por los avances en las técnicas de análisis biológico como por la existencia de amplios estudios clínicos y/o epidemiológicos. Entre las disciplinas y técnicas biomédicas que se consideran asociadas a la biología-medicina de sistemas se incluyen la genómica (estudio de los genes), la transcriptómica (de los transcriptomas), la proteómica (de la estructura y función de las proteínas), la interactómica (de las interacciones moleculares y sus consecuencias), la metabolómica (de las señales moleculares dejadas por los procesos biológicos) y la metagenómica (conjunto de genomas que conviven en un entorno). Es decir, las conocidas popularmente como «omics» u «ómicas». En el campo de la clínica y la epidemiología tenemos diversos ejemplos relacionados con el aparato respiratorio. Unos de los más recientes son los publicados por García-Aymerich et al.[14] y Burgel et al.[15] sobre la tipificación de fenotipos en la EPOC, a partir de un análisis de *cluster* sobre datos de todo tipo procedentes de pacientes reales. Una tercera faceta de la biología-medicina de sistemas, también relacionada con las ciencias «ómicas», es el análisis de los resultados generados por técnicas como los *microarrays* o *chips* de ADN y de proteínas, que se utilizan para analizar simultánea y respectivamente la expresión diferencial de gran número de estos. Son técnicas que han sido ya abundantemente utilizadas en la investigación sobre enfermedades respiratorias. Así, Steiling et al.[16] y Pierrou et al.[17] han demostrado los efectos del tabaco sobre la expresión de múltiples genes relacionados con la lesión celular y el estrés oxidativo en el epitelio bronquial. Sofisticando aún más la complejidad, algunos trabajos reúnen en un mismo análisis datos genéticos con datos clínicos y epidemiológicos[18].

Minería de textos

Como ya se ha mencionado, la minería de textos o datos es un conjunto de técnicas informáticas que permiten el procesamiento de información digital de forma automatizada. Esto permite no solo disponer de dicha información en formato manejable, sino generar nuevo conocimiento. En otros campos del saber biomédico, como la biología molecular o la propia biología de sistemas, este instrumento se ha estado utilizando desde hace años, dando lugar a interesantes resultados[9,19]. En su expresión más sencilla, todos realizamos minería de textos cuando empleamos una herramienta tipo PubMed[20] sobre una base bibliográfica como MEDLINE, seleccionando un tema a partir de palabras clave o autores determinados[21].

Una vez obtenida la información relevante a través de técnicas de minería de textos, se procede al llamado «curado». Esta es una fase que puede realizarse de forma automática, semiautomática o «manual». En el curado automático se pueden añadir requerimientos adicionales, tanto binarios (se escoge o se descarta una información de acuerdo con reglas preestablecidas) como categóricos (p. ej., dando un peso específico al factor de impacto de la revista, al tipo de artículo, al diseño o nivel de evidencia del estudio, al origen de los datos [seres humanos, modelos animales, cultivos celulares], etc.). En el curado manual, un experto en el tema depura la lista de datos o fuentes de información, según su criterio. Sin embargo, y como se discutirá más adelante, con excesiva frecuencia se considera como experto a cualquier profesional de las ciencias biomédicas, no necesariamente ducho en la faceta objeto de la búsqueda.

Un tercer punto importante, una vez seleccionada la información, es establecer conexiones entre datos dispersos. A este proceso se le denomina integración y es fundamental para generar nuevo conocimiento. Resulta obvio que la conexión e integración entre datos obtenidos en diversas áreas de conocimiento no suele producirse de forma espontánea con los métodos tradicionales. Por tanto, se hace necesario que existan métodos automatizados que seleccionen y pongan a disposición del profesional la información potencialmente relevante, independientemente del área del saber en que se haya generado.

A lo largo de los siguientes apartados se detallará cómo puede utilizarse la minería de textos o de datos en las diferentes facetas de la medicina. Entre otras, el seguimiento de hallazgos relevantes en la investigación más básica para su traslado a la clínica, el diagnóstico y la clasificación de la gravedad, y el establecimiento de un pronóstico en algunas de las entidades respiratorias más prevalentes.

Contextualización histórica

Aunque la necesidad de procesar gran cantidad de información de forma automática es relativamente nueva, la tecnología en que se basa la actual minería de textos tiene ya cierto recorrido. Así, hace ya medio siglo se utilizaban técnicas similares para el procesado

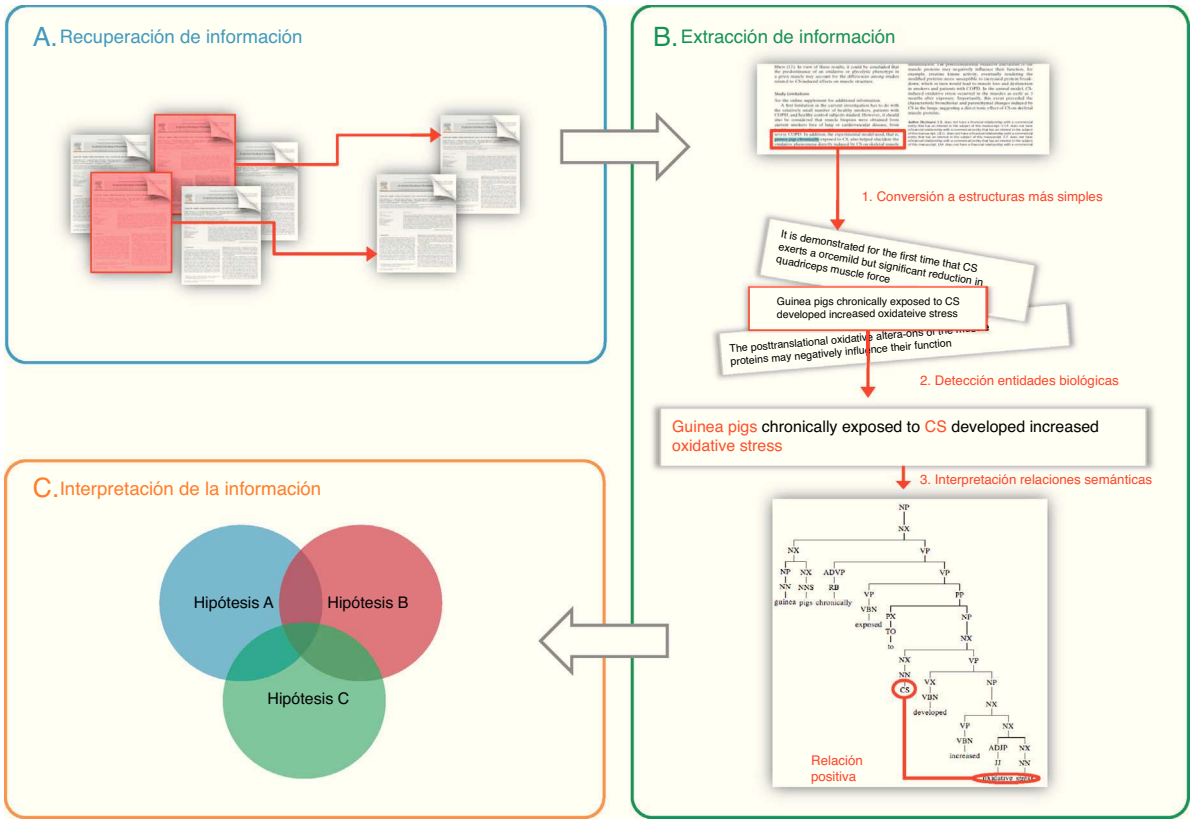

Figura 1. Esquema general de los métodos usados en la *minería de textos* y *de datos:* A) Recuperación de la información; B) Extracción de la información, y C) Interpretación de la información (en ocasiones, la integración de diversas hipótesis previamente contrastadas puede dar lugar a una nueva hipótesis conjunta). En el panel B se ejemplifican los 3 pasos necesarios para extraer la información: 1) descomposición de la información en unidades básicas (p. ej., frases); 2) identificación de las entidades biológicas, y 3) interpretación de las relaciones entre las entidades biológicas.

de discursos[22] o en estudios sobre estructuras lingüísticas[23]. En medicina, algunos de los trabajos pioneros fueron llevados a cabo por Swanson[24,25], que estableció relaciones entre fenómenos aparentemente inconexos, como la migraña y el déficit de magnesio, a partir de los títulos de artículos obtenidos en la base de datos MEDLINE[25]. Estas relaciones fueron validadas experimentalmente años más tarde[26]. Con una metodología parecida, Srinivasan y Libbus[27] y Weeber et al.[28] determinaron el potencial efecto beneficioso de la curcumina y la talidomida en pacientes con enfermedad de Crohn. En la actualidad este tipo de búsquedas se está utilizando ampliamente en medicina y en biología[9] para establecer asociaciones entre, por ejemplo, genes (o proteínas) con enfermedades[29-32], o para valorar interacciones entre diversas proteínas[33,34]. Un instrumento interesante para este objetivo es DisGeNET, diseñado para relacionar la información procedente de las diversas «omics» con la generada sobre las distintas enfermedades[31,32]. En cuanto a la patología respiratoria, y como se verá más adelante, estas técnicas e instrumentos se están utilizando intensivamente para el estudio de diversos aspectos de la EPOC, el asma bronquial y el cáncer de pulmón.

Aspectos generales de la minería de textos

Por definición, la minería de textos y de datos tiene como objetivo recuperar, extraer e interpretar la información almacenada bajo formato electrónico en grandes archivos documentales y bases de datos, mediante el uso de métodos automáticos o semiautomáticos (fig. 1)[9].

La *recuperación de la información* consiste en identificar documentos que contengan datos sobre la pregunta generada. Por ejemplo, ¿cuáles son los biomarcadores que han sido involucrados o presentan potencial para el diagnóstico de la EPOC? En la actualidad, la localización de la información es un problema menor, pues existen infinidad de fuentes accesibles en formato electrónico, tanto originales (revistas, libros de texto, páginas web) como ya recopiladas en bases de datos. Un ejemplo de estas últimas es la ya mencionada MEDLINE, sobre la que actúa el sistema de recuperación de información PubMed[20]. Existen sistemas más específicos, como Textpresso[35,36] o HLungDB[37,38], que es una base de datos sobre cáncer de pulmón que agrupa información sobre genes y proteínas implicados.

El objetivo del segundo paso, la *extracción de información*, consiste en identificar la parte relevante de esta entre todos los datos recuperados. Esencialmente requiere 3 etapas (fig. 1B): a) *procesamiento*, con conversión de textos complejos en estructuras simples (p. ej., palabras o frases cortas), interpretables por los sistemas informáticos; b) *identificación cierta* (estandarizada) de las entidades clínicas o los procesos biológicos a que hacen referencia los documentos, y c) *interpretación* de las relaciones semánticas entre diversas estructuras o entidades. En la extracción de información se utilizan habitualmente sistemas basados en algoritmos de inteligencia artificial[39], métodos estadísticos[40] o sistemas mixtos (p. ej., GENIA Tagger[41,42] o Standford Log-Linear Part-of-Speech Tagger)[43,44].

Un problema inherente a las fases de identificación e interpretación es que los mecanismos automáticos, que permiten filtrar el conocimiento disponible, no poseen el criterio discrecional suficiente. A esto se añade que, en general, los equipos profesionales que diseñan dichos instrumentos carecen de expertos en la materia concreta de la búsqueda. La disociación entre estos últimos y

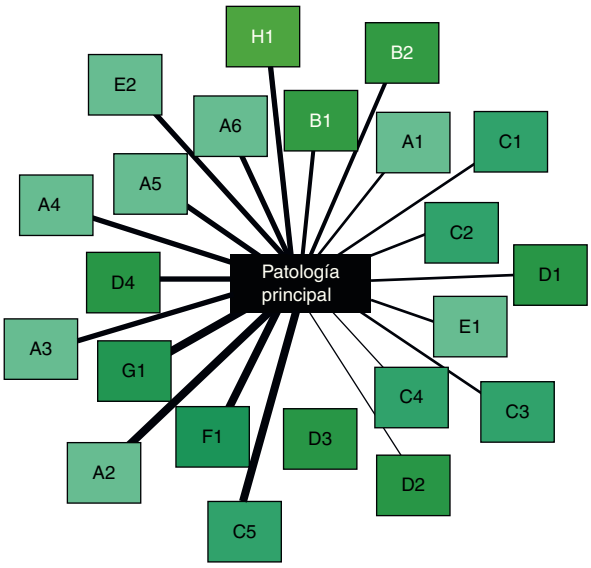

Figura 2. Representación gráfica de las relaciones entre una entidad principal (centro) y diversas comorbilidades potenciales (recuadros de la periferia). El grosor de las líneas entre uno y otros expresa la cantidad de información disponible en cada caso, tras el curado de los datos seleccionados automáticamente. Estas relevancias relativas se han ordenado según rotación horaria. Las comorbilidades de un mismo sistema o aparato del organismo se muestran con un patrón común en el relleno de los recuadros y como letra correlativa de identificación.

los profesionales de la bioinformática es más acusada en el caso de la medicina que en el de la biología, y ha sido identificada como el problema principal en el diseño de los instrumentos de filtrado y curado de la información. Esto es inherente a la parcelación actual de la ciencia en disciplinas, pero se agrava con la progresiva segmentación del conocimiento dentro de cada una. Es pues fundamental incorporar a profesionales clínicos en los equipos multidisciplinares de trabajo[45].

Un problema adicional es que los resultados de los estudios no son axiomas, pero a menudo son tratados como tales por los profesionales procedentes de ciencias más exactas que la medicina. También es importante no olvidar que aunque la minería de datos no es una tecnología totalmente nueva y sus resultados son prometedores, aún existen aspectos que pueden mejorarse. Por ejemplo, a nivel tecnológico existe un problema inherente al lenguaje biomédico: la variabilidad léxica y semántica existente entre las diferentes disciplinas. Es un problema que aunque se puede abordar con cierto éxito mediante algoritmos de inteligencia artificial o métodos estadísticos, todavía no está completamente resuelto. Otro aspecto a tener en cuenta, principalmente en el uso de la minería de textos en medicina, es que las conclusiones obtenidas pueden acabar teniendo efecto sobre los pacientes, con lo que es importante que estén completamente contrastadas.

El último paso es el de la de *interpretación de resultados*. Con él se persigue integrar la información obtenida en los pasos anteriores, y en última instancia obtener asociaciones entre entidades o fenómenos biológicos o clínicos, inicialmente no detectables mediante los métodos convencionales. Volviendo al estudio de Swanson[25], la integración de diversas hipótesis contrastadas individualmente es lo que permitió originar una nueva hipótesis que vinculaba el déficit de magnesio con la migraña. En la figura 2 se muestra como ejemplo una representación gráfica de las asociaciones entre una entidad (la EPOC) y sus comorbilidades. Este tipo de imágenes facilita la interpretación simultánea de información diversa obtenida con la minería de datos, generando nuevas hipótesis.

«*Excavando*» *(mining) informáticamente en el terreno de las enfermedades respiratorias*

Como ya se ha comentado, la alta producción y disponibilidad de datos en el ámbito de la medicina y de las ciencias biológicas más básicas hace de él un espacio apropiado para el uso de la minería de textos. En el campo concreto de las enfermedades respiratorias han aparecido en los últimos años numerosos estudios que hacen uso de esta técnica y que abarcan las diversas etapas que van desde el conocimiento más básico a la medicina aplicada.

Conocimiento en ciencias básicas

En campos como la biología molecular o la fisiología, la minería de textos está originando una transición del modelo científico deductivo clásico —en el que una hipótesis se verifica experimentalmente— a un modelo basado en la búsqueda «a ciegas» de asociaciones entre hechos aparentemente no conectados, pero validables experimentalmente[46,47]. Para ello es conveniente el desarrollo de plataformas especializadas. Así, existen bases de datos unificadas, como la ya mencionada HLungDB[37,38], que han integrado información sobre genes, proteínas, modificaciones epigenéticas y características clínicas, relacionada en todos los casos con el cáncer de pulmón. El principal objetivo de esta plataforma es establecer una red de conexiones basada en moléculas asociadas con esta entidad, facilitando así tanto la investigación más básica como la integración con áreas relativamente distantes como la clínica o la epidemiología. Esto resulta fundamental en entidades como el cáncer de pulmón, que implican alteraciones complejas y de causa multifactorial. Pueden así desarrollarse vías de estudio previamente inexploradas o sugerirse nuevas alternativas terapéuticas. También hay numerosos ejemplos del uso de la minería de textos en aspectos básicos relacionados con la EPOC. Así, Comandini et al.[48] utilizaron datos correspondientes a genes y a expresión de proteínas en sujetos no fumadores, fumadores sin enfermedad pulmonar y fumadores con EPOC, con el objetivo de identificar efectos inducidos por el tabaco. Sus resultados sugieren que determinados genes con actividad antioxidante se sobreexpresan ante la exposición al tabaco en algunos individuos y podrían jugar un papel importante en la protección tisular frente al estrés oxidativo. Estos genes o sus productos podrían ser utilizados como biomarcadores negativos del riesgo para desarrollar una EPOC. En un trabajo muy reciente de algunos miembros de nuestro grupo[49] se ha utilizado la minería de textos para estudiar las comorbilidades de esta misma enfermedad y los mecanismos potencialmente implicados, aflorando la sorprendente abundancia de datos en algunas de las asociaciones (como la de EPOC con cardiopatía isquémica o con cáncer de pulmón) y la escasa presencia en la literatura de otras (como es el caso de las de EPOC con alteraciones nutricionales o con hipertensión pulmonar). También en el caso del asma bronquial se han publicado trabajos interesantes. Por ejemplo, Su et al.[50] han investigado las relaciones de diversos genes entre sí, y de estos con el ambiente en el asma infantil, concluyendo que algunos de ellos condicionan la susceptibilidad a desarrollar asma en los niños. Tremblay et al.[51], por su parte, implementaron una metodología (*Genes to Diseases* [G2D]) basada en la minería de textos para identificar genes candidatos a estar implicados en el desarrollo de asma y atopia. En cuanto a la neumonía y al síndrome del distrés respiratorio del adulto (SDRA) sucede algo parecido. Frenzel et al.[52], analizando múltiples marcadores en el lavado broncoalveolar de pacientes con SDRA, demostraron que la determinación de IL-6 puede ser de gran valor pronóstico.

Parece claro que el análisis automático del conocimiento procedente de diversas ciencias básicas puede aportar nueva luz sobre las enfermedades respiratorias.

Diagnóstico y manejo clínico del paciente respiratorio

Establecer un diagnóstico de certeza y categorizar la gravedad es crucial en cualquier enfermedad. No obstante, esta no es siempre una tarea fácil, sobre todo si el diagnóstico requiere técnicas complejas o que precisan un entrenamiento especial. Este es el caso de la EPOC, donde se necesita tanto la colaboración del paciente como una adecuada maniobra de espirometría forzada. En el momento actual, en algunas iniciativas, como la *Global Initiative for Chronic Obstructive Lung Disease* (GOLD)[53], se requiere además conocer el número de ingresos recientes y la cuantía de la disnea. Utilizar un análisis automático de datos puede permitir verificar la bondad de una determinada maniobra espirométrica, a la vez que aproximar el diagnóstico completo basado en otras variables. Matsumoto et al.[54], mediante minería de datos efectuada en registros electrónicos de unos 27.000 pacientes, observaron que la obstrucción al flujo aéreo detectable por los valores espirométricos no había sido diagnosticada en hasta el 86% de los casos. Este es un ejemplo de algunas aplicaciones de la minería de datos, que permiten además establecer alarmas diagnósticas en las historias clínicas, evitando que hallazgos relevantes pasen desapercibidos al profesional asistencial. Por otra parte, y como se ha visto, el análisis automático de datos puede permitir hallar nuevas relaciones entre variables, generando hipótesis que conduzcan a sistemas alternativos o complementarios de diagnóstico. Respecto del cáncer de pulmón, existen diversos pasos cruciales en su detección y posterior estratificación. Uno de ellos es la valoración de la afectación ganglionar, que se realiza inicialmente mediante técnicas de imagen (tomografía axial computarizada [TAC] y tomografía por emisión de positrones), y se confirma o descarta mediante la obtención de muestras tisulares por ultrasonografía endobronquial o mediastinoscopia. Lu et al.[55] proponen un nuevo método automatizado para la orientación inicial de la afectación ganglionar. Este se basa también en la TAC, pero utilizando luego la minería de datos para comparar los hallazgos con los registros anatómicos de nódulos linfáticos pertenecientes a pacientes ya diagnosticados.

Una fase del proceso diagnóstico que es complementaria a la identificación de la enfermedad es el establecimiento de un pronóstico. Esto implica generalmente la categorización de la gravedad o de la extensión de la enfermedad. En esta fase es igualmente importante disponer de buenos métodos de asignación a cada categoría, pues con frecuencia condicionará el tratamiento. En los últimos años han aparecido sistemas de clasificación automática de las etapas del cáncer de pulmón, que utilizan datos clínicos e histológicos registrados en informes diversos de una o varias instituciones[56,57]. En el caso del trasplante de pulmón o combinado de corazón-pulmón, una adecuada predicción de la supervivencia es crítica no solo para un paciente concreto, sino también para la selección de donantes y receptores. Existe en la actualidad una gran cantidad de información referente a trasplantes (procedimientos, monitorización de pacientes, etc.) que se ha utilizado junto con métodos estadísticos clásicos para realizar predicciones de morbilidad y supervivencia respecto de diversos órganos, incluido el pulmón[58]. Más recientemente se está utilizando también la minería de datos[59], ya que presenta ciertas ventajas sobre los métodos estadísticos clásicos. Por ejemplo, no está limitada por el número de observaciones ni requiere independencia de los observadores, como sucedería en un estudio concreto.

Un campo en el que la minería de datos ha sido particularmente fecunda es el del enfermo respiratorio semicrítico y crítico. >A lo largo de las últimas décadas las unidades a cargo de estos pacientes han implementado sus sistemas de recogida de datos, y como consecuencia directa se han desarrollado 2 aplicaciones fundamentales basadas en la minería de textos.

Por un lado, la generación de modelos predictivos a corto y medio plazo, que permiten crear alertas inteligentes y sistemas de toma de decisiones. Entre los numerosos estudios dedicados al paciente crítico destacan el de Tzavaras et al.[60], que mediante minería de datos y redes neuronales desarrollaron un modelo de soporte a la decisión clínica, basado en la identificación de las variables fisiológicas clave para decidir la instauración de ventilación mecánica.

Por otra parte, se han desarrollado también modelos de largo plazo, utilizados principalmente como método de evaluación de la calidad de los equipos e instituciones implicadas en el cuidado del paciente respiratorio crítico[61-63]. Algunas de estas aplicaciones se han extendido posteriormente a unidades de hospitalización convencional, como las salas de neumología y de cirugía torácica. Tradicionalmente, estos instrumentos se han construido utilizando únicamente datos demográficos y administrativos, para determinar índices de supervivencia, de mortalidad, etc. Sin embargo, de forma más reciente, se están aprovechando los datos fisiológicos y clínicos que quedan almacenados para generar modelos basados en la minería de datos. Existen diversos estudios, como los de Bohensky et al.[62] o Kim et al.[63], que han demostrado que estos modelos son superiores a los métodos clásicos en la predicción de supervivencia.

Las *vías clínicas* son el registro de los procedimientos realizados sobre un paciente concreto durante su estancia hospitalaria. También en este campo se han publicado estudios donde se emplean métodos de análisis automatizado de datos[64,65]. Las vías clínicas constituyen una herramienta básica de gestión de la calidad asistencial y se ha demostrado que su implementación permite disminuir la variabilidad en la práctica clínica. Otras ventajas son la ayuda al clínico en la toma de decisiones y una optimización de los recursos empleados, con reducción del tiempo y costes en la atención.

Conclusiones

Se dispone hoy en día de gran cantidad de información científica y médica. No obstante, existe un problema inherente a tal volumen de datos: su recuperación selectiva e interpretación se hacen prácticamente imposibles para un profesional si emplea los métodos clásicos. En ese escenario, el uso de herramientas bioinformáticas como la minería de textos o datos adquiere una relevancia fundamental. Estas herramientas, que han tenido ya un papel importante en otros campos del saber biomédico, se han empezado a utilizar recientemente en la medicina respiratoria. Los niveles más evidentes de aplicación médica (en investigación y en clínica) de la minería de textos son la integración y la transferencia de los avances obtenidos en las ciencias más básicas, y una mejor comprensión de los procesos de diagnóstico, de categorización de la gravedad y de establecimiento de pronóstico de las enfermedades. También puede ser de gran utilidad para generar modelos predictivos de *outcomes*, crear alertas inteligentes y ayudar al clínico en la toma de decisiones.

Financiación

CB06/06/0043 (CIBERES), SAF2011-26908 (Ministerio de Economía y Competitividad) y 2009SGR393 (Generalitat de Catalunya).

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Agradecimientos

Deseamos agradecer al Dr. Ferran Sanz y la Dra. Laura Furlong, del Programa de Investigación en Informática (GRIB, IMIM-UPF),

por su asesoramiento y consejos en la redacción del presente manuscrito.

Bibliografía

1. World Health Organization (2005) WHO global report: Preventing chronic diseases: A vital investment. Geneva: World Health Organization [consultado Mar 2013]. Disponible en: http://www.who.int/chp/chronic_disease_report/en
2. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006;3:e442.
3. Rosenbaum L, Lamas D. Facing a slow-motion disaster— the UN meeting on noncommunicable diseases. N Engl J Med. 2011;365:2345–8.
4. Buist AS, McBurnie MA, Vollmer WM, Gillespie S, Burney P, Mannino DM, et al., BOLD Collaborative Research Group. International variation in the prevalence of COPD (the BOLD Study): A population-based prevalence study. Lancet. 2007;370:741–50.
5. Mannino DM, Buist AS. Global burden of COPD: Risk factors, prevalence, and future trends. Lancet. 2007;370:765–73.
6. Instituto Nacional de Estadística [consultado Mar 2013]. Disponible en: http://www.ine.es
7. Divo M, Cote C, de Torres JP, Casanova C, Marin JM, Pinto-Plata V, et al., BODE Collaborative Group. Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2012;186:155–61.
8. Meyfroidt G, Güiza F, Ramon J, Bruynooghe M. Machine learning techniques to examine large patient databases. Best Pract Res Clin Anaesthesiol. 2009;23:127–43.
9. Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug Discov Today. 2005;10:439–45.
10. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000:270–4.
11. UpToDate [consultado Mar 2013]. Disponible en: http://www.uptodate.com
12. Noble D. The Music of Life: Biology Beyond the Genome. Oxford: Oxford University Press; 2006. p. 21.
13. Sobradillo P, Pozo F, Agustí A. P4 medicine: The future around the corner. Arch Bronconeumol. 2011;47:35–40.
14. Garcia-Aymerich J, Gómez FP, Benet M, Farrero E, Basagaña X, Gayete A, et al., PAC-COPD Study Group. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. Thorax. 2011;66:430–7.
15. Burgel PR, Paillasseur JL, Peene B, Dusser D, Roche N, Coolen J, et al. Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality. PLoS One. 2012;7:e51048.
16. Steiling K, van den Berge M, Hijazi K, Florido R, Campbell J, Liu G, et al. A dynamic bronchial airway gene expression signature of COPD and lung function impairment. Am J Respir Crit Care Med. 2013;187:933–42.
17. Pierrou S, Broberg P, O'Donnell RA, Pawłowski K, Virtala R, Lindqvist E, et al. Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2007;175:577–86.
18. Siedlinski M, Tingley D, Lipman PJ, Cho MH, Litonjua AA, Sparrow D, et al., and the COPDGene and ECLIPSE Investigators. Dissecting direct and indirect genetic effects on chronic obstructive pulmonary disease (COPD) susceptibility. Hum Genet. 2013;132:431–41.
19. Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol. 2006;24:571–9.
20. PubMed NCBI [consultado Mar 2013]. Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/
21. De Granda-Orive JI, Alonso-Arroyo A, Villanueva Serrano SJ, Aleixandre-Benavent R, González-Alcaide G, García-Río F, et al. Comparison between two five year periods (1998/2002 and 2003/2007) on the production, impact and co-authorship of publications on tobacco and smoking by Spanish authors using the Science Citation Index. Arch Bronconeumol. 2011;47:25–34.
22. Harris Z. Discourse analysis. Language. 1952;28:18–23.
23. Harris Z. Co-occurrence and transformation in linguistic structure. Language. 1957;33:283–340.
24. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30:7–18.
25. Swanson DR. Migraine and magnesium: Eleven neglected connections. Perspect Biol Med. 1988;31:526–57.
26. Ramadan NM, Halvorson H, Vande-Linde A, Levine SR, Helpern JA, Welch KM. Low brain magnesium in migraine. Headache. 1989;29:416–9, and 590-593.
27. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics. 2004;20 Suppl 1:i290–6.
28. Weeber M, Vos R, Klein H, de Jong-van den Berg LT, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inform Assoc. 2003;10:252–9.
29. Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, et al. Systematic association of genes to phenotypes by genome and literature mining. PLoS Biol. 2005;3:e134.
30. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, et al. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pac Symp Biocomput. 2006:4–15.
31. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. DisGeNET: A Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics. 2010;26:2924–6.
32. DisGeNET [consultado Mar 2013]. Disponible en: http://ibi.imim.es/DisGeNET/web/v02/home/
33. Hao Y, Zhu X, Huang M, Li M. Discovering patterns to extract protein-protein interactions from the literature: Part II. Bioinformatics. 2005;21:3294–300.
34. Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein-protein interactions. Bioinformatics. 2001;17:359–63.
35. Muller HM, Kenny EE, Sternberg PW. Textpresso: An ontology-based information retrieval and extraction system for biological literature. PLoS Biology. 2004;2:e309.
36. Textpresso [consultado Mar 2013]. Disponible en: http://www.textpresso.org
37. Wang L, Xiong Y, Sun Y, Fang Z, Li L, Ji H, et al. HLungDB: An integrated database of human lung cancer research. Nucleic Acids Res. 2010;38:D659–65.
38. HLungDB [consultado Mar 2013]. Disponible en: http://www.megabionet.org/bio/hlung/
39. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: A machine learning approach. Bioinformatics. 2001;17 Suppl 1:S97–106.
40. Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics. 2006;22:3089–95.
41. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus-semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19 Suppl 1:i180–2.
42. GENIA Tagger [consultado Mar 2013]. Disponible en: http://www.nactem.ac.uk/tsujii/GENIA/tagger
43. Toutanova K, Klein D, Manning CD, Singer Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. En: Proceedings of HLT-NAACL. 2003. p. 252–9.
44. Standford Log-Linear Part-of-Speech Tagger [consultado Mar 2013]. Disponible en: http://nlp.stanford.edu/software/tagger.shtml
45. Cases M, Furlong LI, Albanell J, Altman RB, Bellazzi R, Boyer S, et al. How to improve data and knowledge management to better integrate healthcare and research. J Int Med. 2013;274:321–8.
46. Brent R, Lok L. Cell biology. A fishing buddy for hypothesis generators. Science. 2005;308:504–6.
47. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays. 2004;26:99–105.
48. Comandini A, Marzano V, Curradi G, Federici G, Urbani A, Saltini C. Markers of anti-oxidant response in tobacco smoke exposed subjects: A data-mining review. Pulm Pharmacol Ther. 2010;23:482–92.
49. Grosdidier S, Ferrer A, Faner R, Gea J, Piñero J, Roca J, et al. Exploring the Diseasome of COPD and its associated diseases. Virtual Physiological Human Network of Excellence (VPH NoE) 2 nd Conference. Londres (UK); 2012: Abstract no. 42.
50. Su MW, Tung KY, Liang PH, Tsai CH, Kuo NW, Lee YL. Gene-gene and gene-environmental interactions of childhood asthma: A multifactor dimension reduction approach. PLoS One. 2012;7:e30694.
51. Tremblay K, Lemire M, Potvin C, Tremblay A, Hunninghake GM, Raby BA, et al. Genes to diseases (G2D) computational method to identify asthma candidate genes. PLoS One. 2008;3:e2907.
52. Frenzel J, Gessner C, Sandvoss T, Hammerschmidt S, Schellenberger W, Sack U, et al. Outcome prediction in pneumonia induced ALI/ARDS by clinical features and peptide patterns of BALF determined by mass spectrometry. PLoS One. 2011;6:e25544.
53. Global Initiative for Chronic Obstructive Lung Disease (GOLD) [consultado Mar 2013]. Disponible en: http://www.goldcopd.org/
54. Matsumoto K, Takahashi Y, Gon Y, Akahoshi T, Nakayama T, Asai S, et al. Identifying unrecognized airflow obstruction in cases with lifestyle-related diseases using a data mining system with electronic medical records. Rinsho Byori. 2011;59:128–33.
55. Lu K, Taeprasartsit P, Bascom R, Mahraj RP, Higgins WE. Automatic definition of the central-chest lymph-node stations. Int J Comput Assist Radiol Surg. 2011;6:539–55.
56. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc. 2010;17:440–5.
57. Nguyen A, Moore D, McCowan I, Courage MJ. Multi-class classification of cancer stages from free-text histology reports using support vector machines. Conf Proc IEEE Eng Med Biol Soc. 2007;2007:5140–3.
58. Lin HM, Kauffman HM, McBride MA, Davies DB, Rosendale JD, Smith CM, et al. Center-specific graft and patient survival rates: 1997 United Network for Organ Sharing (UNOS) report. JAMA. 1998;280:1153–60.
59. Oztekin A, Denle D, Kong ZJ. Predicting the graft survival for heart-lung transplantation patients: An integrated data mining methodology. Int J Med Inform. 2009;78:e84–96.
60. Tzavaras A, Weller PR, Prinianakis G, Lahana A, Afentoulidis P, Spyropoulos B. Locating of the required key-variables to be employed in a ventilation management decision support system. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:112–5.

61. Saeed M, Mark R. A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations. AMIA Annu Symp Proc. 2006;2006:679–83.
62. Bohensky MA, Jolley D, Pilcher DV, Sundararajan V, Evans S, Brand CA. Prognostic models based on administrative data alone inadequately predict the survival outcomes for critically ill patients at 180 days post-hospital discharge. J Crit Care. 2012;27:e11–21.
63. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res. 2011;17:232–43.
64. Huang Z, Lu X, Duan H, Fan W. Summarizing clinical pathways from event logs. J Biomed Inform. 2013;46:111–27.
65. Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med. 2012;56:35–50.