Gac Sanit. 2016;30(1):63–65

Debate

Big data en sanidad en España: la oportunidad de una estrategia nacional

Big data in health in Spain: now is the time for a national strategy

Carlos Luis Parra Calderón

Grupo de Investigación e Innovación en Informática Biomédica, Ingeniería Biomédica y Economía de la Salud, Instituto de Biomedicina de Sevilla-Hospital Universitario Virgen del Rocío, Sevilla, España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 14 de julio de 2015
Aceptado el 13 de octubre de 2015
On-line el 21 de noviembre de 2015

Aplicaciones de los *big data* en la investigación en salud pública

Existe ya una gran base de conocimiento sobre los resultados de la aplicación de los *big data* en salud y biomedicina, incluso en salud pública en particular. Conforme aumenta su aplicación se identifican nuevos retos a los que enfrentarse, así como nuevas oportunidades que acrecientan el interés por el desarrollo de la investigación en este dominio.

Los métodos y herramientas de *big data* se caracterizan por el volumen, la complejidad y la velocidad de la información que manejan. Las propiedades, los retos y los asuntos relevantes que caracterizan la aplicación de los *big data* en biomedicina son la gran variedad en la naturaleza de los datos y la alta velocidad de proceso requerida; retos relacionados con la veracidad de los datos, con los flujo de trabajo, con los métodos computacionales, con la extracción de información significativa, con el intercambio de datos y con la necesidad de expertos en el uso de estas tecnologías. Son relevantes asuntos relacionados con la reutilización de datos, con el riesgo de falso descubrimiento de conocimiento y con la privacidad. La propia definición de *big data* en salud tiene sus matices[1].

En Internet, la información sobre las enfermedades y sus brotes se difunde no sólo a través de noticias de las agencias de los gobiernos, sino también por canales informales, que van desde la prensa a los blogs, mediante análisis de los registros de búsquedas en la web. Las diversas e inmensas fuentes de información proporcionan una vista de la salud global diferente de la que se deriva de las infraestructuras de salud pública tradicionales[2].

Una publicación pionera se centró en la localización de registros de consultas de Google para detectar la actividad de la gripe en regiones específicas con grandes poblaciones de usuarios que hacen búsquedas en las webs[3], y desde entonces han aparecido muchas extensiones de este estudio inicial.

Por otra parte, los *big data* proporcionan un complemento útil en la medida en que se han convertido en un componente importante para la vigilancia de enfermedades infecciosas como la gripe[4].

Más allá de la información disponible en Internet, los repositorios de información de salud pública se encuentran en transición hacia centros de datos centralizados[5]. Los enfoques *big data* permiten incorporar capacidades de geolocalización obtenidas por la dirección de los ciudadanos, así como por la dirección de hospitales, farmacias, médicos y ambulatorios, lo que hace posible obtener mapas geográficos de salud en el tiempo para toda la población de una región. Una aplicación de futuro para los *big data* en salud pública es la integración de otras fuentes de información, como son contaminantes, tráfico, calefacción, tiendas de comestibles y mercados, o insumos alimenticios, que podrían mejorar la precisión de la estratificación por riesgo de la población[6].

El desarrollo de estrategias nacionales de informática sanitaria permite disponer de una gran fuente integrada de información que ofrece una imagen completa de la salud de una determinada región, como es el ejemplo de Dinamarca[7]. De hecho, en el ámbito de la vigilancia y la intervención en salud pública, destaca la aplicación de los *big data* a la información disponible en grandes bases de datos de historias de salud electrónicas, teniendo en cuenta las posibles limitaciones de una inferencia correcta de causa-efecto.

En este sentido, se demuestra que el uso de *big data* puede ofrecer oportunidades para reducir costes en el tratamiento de la información clínica de las historias de salud electrónicas, tal como se demuestra en el estudio con seis casos de pacientes de alto coste[8] en los ámbitos de los reingresos, *triage*, descompensaciones y eventos adversos, y en el tratamiento de enfermedades que afectan a múltiples órganos y sistemas. En este sentido, los *big data* aplicados a la información clínica harán que esté disponible una nueva generación de herramientas más inteligentes de soporte a la decisión clínica guiada por datos en tiempo real[9].

En el ámbito de la investigación traslacional se están desarrollando de manera intensa nuevas aplicaciones de tecnologías *big data*, como *hadoop* en secuenciación NGS *(Next Generation Sequencing)*, así como en la fenotipificación de pacientes basándose en la información de la historia de salud electrónica. Muy relevante es la experiencia del consorcio eMERGE aportando un marco de

Correo electrónico: carlos.parra.sspa@juntadeandalucia.es

http://dx.doi.org/10.1016/j.gaceta.2015.10.005

experiencia fundamental en la asociación genotipo-fenotipo para el descubrimiento genómico y para validar nuevos estándares de representación y normalización de la información, cuyas características fundamentales son la heterogeneidad y la complejidad en estos dominios[10,11].

Por último, son prometedores los avances en herramientas de análisis visual en salud pública[12].

Aportaciones de valor en el establecimiento de relaciones causales, tanto en investigación etiológica como en investigación evaluativa

Estas técnicas se están aplicando ya con éxito para el descubrimiento de factores de riesgo y de estudios genotipo-fenotipo. Sin embargo, es importante tener en cuenta el rigor en el uso de los términos «asociación» respecto a «causalidad»[13]. Al igual que con la vigilancia de enfermedades, también se ha demostrado que los métodos *big data* proporcionan información valiosa acerca de los eventos adversos de los medicamentos, en particular las reacciones causadas por combinaciones específicas de estos.

Un reto fundamental es la extracción de información de los textos narrativos en la historia de salud electrónica. Mediante el desarrollo de métodos para extraer y hacer uso de estos complejos relatos clínicos a gran escala, será posible un análisis matizado de la salud del paciente a través de la historia de salud electrónica, y finalmente se formará una imagen más completa de complejos conjuntos de características que influyen en el diagnóstico y el tratamiento de las enfermedades[14]. Esto requiere validaciones robustas de la aplicación de los algoritmos en este sentido, como es el caso del procesamiento del lenguaje natural para estimar la prevalencia y la gravedad de las enfermedades a partir de su aplicación en la historia de salud electrónica en Nueva Zelanda[15].

Para dar soporte a estos retos se perfilan nuevos perfiles profesionales, como los *biocurators* o *data managers* especializados en datos de naturaleza biológica, que deben desarrollar capacidades en el ámbito de los métodos y las herramientas de *big data*[16].

El modelado de datos para su tratamiento en *big data* a menudo puede conducir a una correlación o inferencia estadística sesgada, lo que se conoce como «falso descubrimiento». Usuarios de *big data* clínicos se enfrentan a retos importantes ya conocidos, pero con una dimensión desconocida hasta ahora, como son el tamaño de la muestra, el sesgo de selección, el problema de la interpretación, los valores perdidos, problemas de dependencia y metodologías de manejo de datos adecuadas[13].

Es fundamental plantear soluciones adecuadas a los retos específicos de análisis en función de la naturaleza de los datos: imagen médica, señales biomédicas e información genómica integrada con información fisiológica[17]. En este sentido se está trabajando en la infraestructura de *MapReduce* en plataformas *Hadoop*.

Por último, es relevante destacar las experiencias de infraestructura de *software* libre y orientadas a servicios, de análisis basado en *big data*, para mejorar el uso del acceso a datos heterogéneos y de fuentes diversas, como es el proyecto *SOCR Data Dashboard*.

La oportunidad de implementar una estrategia nacional de *big data* en España

Los retos a los que nos enfrentamos y que hemos planteado no pueden demorar más una respuesta ordenada en el Sistema Nacional de Salud que potencie los efectos beneficiosos de la aplicación de *big data* en sanidad y en biomedicina en España, reduciendo riesgos como pueden ser la pérdida de economías de escala en las inversiones tecnológicas requeridas o la dificultad de escalabilidad y de explotación unificada si las iniciativas no están alineadas. La tecnología está disponible y la industria que la ofrece muestra una agresividad comercial muy alta para introducirla (a cualquier precio), influenciada por la travesía del desierto en la que se encuentra desde que se iniciaron la crisis económica y los consiguientes dramáticos recortes en la inversión en tecnologías de la información y la comunicación. Estamos frente a una oportunidad histórica para aunar voluntades, políticas y tecnologías en una estrategia nacional. En este sentido, es procedente desarrollar una estrategia inicial en el ámbito de la investigación biomédica, donde los retos son tremendos pero los posibles beneficios de una explotación masiva de la información digital disponible en el Sistema Nacional de Salud son evidentes, alineando esfuerzos de las comunidades autónomas. Destaca como referencia en este país y primera experiencia importante de *big data* sobre información de pacientes el proyecto *Visc+* de Cataluña, que está comenzando su recorrido para uso científico.

Ejemplos internacionales relevantes son la iniciativa *care.data* del National Health Service del Reino Unido para la apertura del acceso a los registros clínicos para la investigación[18], o la estrategia *Big Data to Knowledge BD2 K* de infraestructura de los National Institutes of Health en los Estados Unidos[19].

Conclusiones

La aplicación de *big data* en sanidad es imparable, y ya existen referencias suficientes para conocer sus limitaciones y riesgos frente a los posibles beneficios que ofrece. En la agenda digital de España procede el desarrollo de una estrategia nacional en la que se tengan en cuenta todos estos factores, priorizando la implementación de casos de uso de valor compartido e incorporando un marco claro y viable de medición del impacto de dicha estrategia.

Contribuciones de autoría

C.L. Parra es el único autor.

Financiación

Ninguna.

Conflictos de intereses

Ninguno.

Bibliografía

1. Baro E, Degoul S, Beuscart R, et al. Toward a literature-driven definition of big data in healthcare. BioMed Research International. 2015;2015:639021.
2. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection — harnessing the web for public health surveillance. N Engl J Med. 2009;360:2153–7.
3. Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4.
4. Milinovich GJ, Williams GM, Clements AC, et al. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis. 2014;14:160–8.
5. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearbook of Medical Informatics. 2014;9:8-13.
6. Martin Sanchez F, Gray K, Bellazzi R, et al. Exposome informatics: considerations for the design of future biomedical research information systems. J Am Med Inform Assoc. 2014;21:386–90.
7. Sortsø C, Thygesen LC, Brønnum-Hansen H. Database on Danish population-based registers for public health and welfare research. Scand J Public Health. 2011;39:17–9.
8. Bates DW, Saria S, Ohno-Machado L, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood). 2014;33:1123–31.
9. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309:1351–2.
10. Chute CG, Ullman-Cullere M, Wood GM, et al. Some experiences and opportunities for big data in translational research. Genet Med. 2013;15:802–9.
11. Gottesman O, Kuivaniemi H, Tromp G, et al. eMERGE network. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. Genet Med. 2013;15:761–71.

12. Ola O, Sedig K. The challenge of big data in public health: an opportunity for visual analytics. Online Journal of Public Health Informatics. 2014;5:223.
13. Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. Eysenbach G, editor. JMIR Medical Informatics. 2014;2:e1.
14. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearbook of Medical Informatics. 2014;9:14–20.
15. MacRae J, Darlow B, McBain L, et al. Accessing primary care big data: the development of a software algorithm to explore the rich content of consultation records. BMJ Open. 2015;5:e008160.
16. Howe D, Costanzo M, Fey P, et al. Big data: the future of biocuration. Nature. 2008;455:47–50.
17. Belle A, Thiagarajan R, Soroushmehr SM, et al. Big data analytics in healthcare. Biomed Res Int. 2015:370194.
18. The care.data programme–collecting information for the health of the nation; NHS. Disponible en: http://www.england.nhs.uk/ourwork/tsd/care-data/
19. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2 K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21:957–8.