Gac Sanit. 2016;30(1):66–68

Debate

Peligros del uso de los *big data* en la investigación en salud pública y en epidemiología

Risks of the use of big data in research in public health and epidemiology

Glòria Pérez

Agència de Salut Pública de Barcelona, Barcelona, España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 13 de julio de 2015
Aceptado el 30 de septiembre de 2015
On-line el 17 de noviembre de 2015

Even John Snow needed to start with a plausible hypothesis to know where to look and choose what data to examine[1].

La realidad incuestionable es la aparición de los *big data* (datos masivos). Este término se refiere a los grandes volúmenes de información compleja y conectable que crece continuamente, de modo que la información parece duplicarse cada 2 años, y este fenómeno podría estarse acelerando. En este sentido, cabe destacar que mucha de esta información era inaccesible hace solo una década.

Los datos masivos proceden de múltiples fuentes de información, derivados de diferentes contextos, tales como los financieros, la informática de negocio, el ocio, las redes sociales y las redes laborales, las ciencias ambientales y también la salud. En este último ámbito existen múltiples fuentes de información derivadas de la medicina asistencial, la genómica, la biología molecular, la clínica, la epidemiología y la salud pública, entre otras.

La investigación en salud pública y en epidemiología tiene por objetivo conocer la salud de la población y sus determinantes[2]. Los posibles beneficios de los *big data* en la investigación en este campo son el uso de diversas fuentes de información y la rapidez en el análisis[3]. Estas dos características, según algunas opiniones, podrían dejar el método científico actual obsoleto[4]. No comparto esta última opinión. Parece que nos volvemos a enfrentar al mismo problema que hace tres décadas con la llegada de los ordenadores personales, cuando se creía que la velocidad de análisis iba a cambiar el método científico en la investigación epidemiológica. Es por ello que centraré mi contribución a este debate en señalar los «peligros» del uso de los *big data* en la investigación en salud pública y en epidemiología.

La necesidad de hipótesis

Disponer de datos es una de las bases para el progreso científico. En investigación usamos modelos, a veces complejos, como una forma de aproximación a la realidad. Estos modelos de análisis de datos se sustentan en hipótesis y en marcos conceptuales, sin los cuales sería imposible realizar investigación. Aunque parece claro que las hipótesis han de guiar la investigación cualquiera que sea el volumen de datos, existen diversas posiciones al respecto. Por un lado, están las personas que creen que los datos nos dirán aquello que queremos saber. Esta posición es muy cercana al «ir de pesca» en los datos, adjudicándoles un cierto «buenismo» debido a que el gran tamaño nos permitirá realizar inferencias estadísticas fiables[4]. En el otro extremo se situarían aquellas personas que creen que analizar los *big data* es analizar terabytes de ruido para obtener un megabyte de señal, y por tanto usarían los *big data* en modelos causales más o menos simples que se prueban en entornos muy controlados.

Estas dos posiciones están explicadas de una forma un tanto simplista, pero describen las dificultades con que nos enfrentamos las personas que nos dedicamos a la investigación, sin que por el momento tengamos una comprensión demasiado sólida de cómo abordar de manera sistemática y eficiente lo que suponen los *big data* en la investigación en salud pública y en epidemiología[5].

El origen de los *big data* y sus posibles sesgos

Los datos útiles para la investigación en salud pública y en epidemiología proceden habitualmente de fuentes diseñadas ad hoc para la investigación o bien de fuentes secundarias, como las historias clínicas, pruebas de laboratorio, censo de población, registros de enfermedades, etc. Lo que distinguiría al entorno *big data* es, por un lado, la incorporación de otras fuentes de información, como las derivados de los servicios prestados por las App de e-salud, *wereables*, las redes sociales o las plataformas «nube», entre otras, y la posibilidad de realizar la consulta a múltiples fuentes de datos *online*[4].

Hay que señalar que los datos que se obtienen de estas plataformas son muestras de conveniencia y pueden tener un número importante de sesgos de selección y de información, de los cuales no nos protege el tamaño de los datos. Un ejemplo de sesgo de información podría ser el uso de los Twitterbots, programas usados para producir mensajes automatizados que permiten, mediante el acceso a potenciales clientes, mejorar el posicionamiento de una empresa. Al contrario, puede surgir un sinnúmero de asociaciones,

Correo electrónico: gperez@aspb.cat

http://dx.doi.org/10.1016/j.gaceta.2015.09.007

algunas de ellas debidas al azar y a la existencia de sesgos como el de confusión. También, las empresas de estas plataformas mejoran los servicios a los usuarios constantemente, lo cual podría afectar a la comparabilidad de los datos a lo largo del tiempo. Tampoco es fácil obtener datos y replicar los resultados de los estudios para poder determinar su robustez.

El análisis de los datos

La minería de datos es la exploración automática o semiautomática de los grandes conjuntos de datos con la intención de descubrir patrones. Es uno de los pasos que componen el proceso del *knowledge discovery in databases*[6], en el cual se incluyen la recolección y la preparación de los datos, la interpretación de los resultados y la información de estos. Sin embargo, la minería de datos genera ciertos desafíos para la ciencia actual[7]. El primero, como ya se ha mencionado, es la búsqueda de patrones en los *big data.* Para ilustrarlo, Shiffrin[7] pone un ejemplo: «Supongamos una base de datos de un terabyte de datos con la posibilidad de contener mil factores medibles. El número de posibles correlaciones de esos factores sería del orden de dos por mil». Y el segundo es la posibilidad de asociaciones espurias, que Shiffrin[7] expone claramente: «En una base de terabytes de datos, el factor A se correlaciona con el factor B, y esta podría ser una relación causal directa entre ambos factores; sin embargo, también podría haber unos 10^{310} otros potenciales bucles causales y las distribuciones de probabilidad asignadas a las 10^{310} posibilidades».

La tecnología permite y permitirá analizar un ingente volumen de datos y establecer innumerables asociaciones mediante modelos complejos. Habrá que desarrollar nuevas propuestas que traten los niveles de significación estadística de forma diferente, tal como se hizo evidente al tener que desarrollar los *Manhattan plot*[8] para los estudios de epidemiología genética. Sin embargo, la mayor complejidad de las herramientas analíticas podría tener como consecuencia posibles limitaciones en la transparencia de los métodos y en la interpretación y la replicabilidad de los resultados[9].

Todo ello nos lleva a recordar los criterios de causalidad de Bradford Hill, en los que la fuerza de asociación estadística es solo una de las nueve condiciones para establecer la causalidad[10].

La generación de conocimiento y su transferencia

La generación de conocimiento es un proceso dinámico de síntesis, interpretación, integración y difusión de los resultados de la investigación[11]. Es indudable que Internet ha permitido la mejora del trabajo de campo de las encuestas, la recogida de datos y los procesos de compartir datos y de intercambio del conocimiento[12], como ya está ocurriendo en algunas redes internacionales (por ejemplo, la de la malaria[13] y la de demografía[14]).

Sin embargo, existen otros ámbitos, como son las predicciones de alertas con consecuencias para la salud de la población, en los cuales, aunque se ha demostrado el alto valor alcanzado, aún se está lejos de poder suplantar a los métodos más tradicionales[15].

Tampoco puede desecharse la posibilidad de la manipulación por parte de empresas con ánimo de lucro, o bien desde visiones corporativas que muy lícitamente para sus intereses pretenden influir mediante los *big data* en las decisiones sobre la salud de la población, sin obviar que puedan tener una mayor capacidad de transferencia que las instituciones públicas encargadas de la salud pública.

Un aspecto no desdeñable es poder refutar o aceptar resultados de estudios basados en los *big data*. No obstante, requerirá que el estudio esté bien sustentado metodológicamente, sea cual sea el origen de los datos.

Aspectos sociales, éticos y políticos de la investigación con *big data*

La regulación europea prevé la protección de los datos personales, entre los que se encuentran los de la salud de la ciudadanía[16]. Sin embargo, existen países donde la normativa puede ser más laxa o inexistente, y donde obtener estos datos puede ser más fácil. Por otro lado, la dependencia económica de los países de renta baja imposibilita que ejerzan la soberanía sobre sus datos frente a los países de renta alta. A lo anterior cabría añadir que en la mayoría de los casos es difícil que los resultados de las investigaciones reviertan en la población que los ha originado, debido a la inestabilidad política, la corrupción, la pobreza y la precariedad de los sistemas de salud y del acceso a las nuevas tecnologías. Los avances científicos que se deriven de esas investigaciones deberían mejorar la salud y los determinantes de la salud de la población en esos países.

Conclusiones y recomendaciones

Se concluye que las buenas prácticas en la investigación en salud pública y en epidemiología no han de ser diferentes para las investigaciones que usen *big data.* Por tanto, la división entre la investigación con *big data* y la investigación tradicional no parece pertinente.

Los investigadores e investigadoras de la salud pública y la epidemiología deberían desempeñar un papel central en la propuesta de hipótesis innovadoras, en la construcción de infraestructuras para el almacenamiento de grandes conjuntos de datos y en asegurar el desarrollo de enfoques sistemáticos en el análisis de grandes conjuntos de datos complejos y masivos. Para ello, las sociedades científicas relacionadas con la salud pública y la epidemiología deberían proponer una estrategia formativa y abrir un debate necesario en nuestro colectivo.

Contribuciones de autoría

Autora única.

Conflicto de intereses

Parte de este texto se presentó como comunicación oral al II Congreso Iberoamericano de Epidemiología y Salud Pública.

La autora declara que pertenece al comité editorial de Gaceta Sanitaria, pero que no ha participado en el proceso editorial del manuscrito.

Bibliografía

1. Khoury BMJ, Ioannidis JPA. Big data meets public health. Science. 2014;346:1054–5.
2. Chun-Hai-Fung I, Tsz-Ho-Tse Z, Fu K-W. Converting big data into public health. Science. 2015;347:620.
3. Harvard School of Public Health. Big data's big visionary. Magazine. [Internet]. Harvard; 2014. p. 32-49. (Consultado el 10/05/2015.) Disponible en: http://www.hsph.harvard.edu/news/magazine/big-datas-big-visionary/
4. Standen A. How big data is changing medicine listen: KQED Science [Internet]. 2014. (Consultado el 10/05/2015.) Disponible en: http://ww2.kqed.org/science/2014/09/29/how-big-data-is-changing-medicine/
5. Birney E. The making of ENCODE: lessons for big-data projects. Nature [Internet]. 2012;489:49–51 (Consultado el 10/05/2015.) Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/22955613
6. Fayyad U, Piatetsky-shapiro G, Smyth P. From data mining to knowledge discovery in. Intell Artif Mag. 1996;17:37–54.
7. Shiffrin R. Introduction to the Sackler Colloquium, drawing causal inference from big data. En: Introduction to Sackler Colloqium [Internet]. Washington, D.C.: National Academy of Sciences; 2015 (Consultado el 10/05/2015.) Disponible en: http://www.nasonline.org/programs/sackler-colloquia/completed_colloquia/Big-data.html?referrer=https://www.google.es/

8. Gibson G. Hints of hidden heritability in GWAS. Nat Genet [Internet]. Nature Publishing Group;. 2010;42:558–60 (Consultado el 10/05/2015.) Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/20581876
9. Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Information, Commun Soc. 2012;15:662–79.
10. Hill A-B. President's address the environment and disease. Proc R Soc Med. 1965;58:295–300.
11. Khoury MJ, Lam TK, Ioannidis JP, et al. Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiol Biomarkers Prev [Internet]. 2013;22:508–16 (Consultado el 10/05/2015.) Disponible en: http://cebp.aacrjournals.org/cgi/doi/10.1158/1055-9965.EPI-13-0146
12. Lang T. Advancing global health research through digital technology and sharing data. Science. 2011;331:714–7.
13. Hay SI, Snow RW. The Malaria Atlas Project: developing global maps of malaria risk. PLoS Med [Internet]. 2006;3:e473 (Consultado el 10/05/2015.) Disponible en: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1762059&tool=pmcentrez&rendertype=Abstract
14. Kowal P, Kahn K, Ng N, et al. Ageing and adult health status in eight lower-income countries: the INDEPTH WHO-SAGE collaboration. Glob Health Action [Internet]. 2010;3:11–22 (Consultado el 10/05/2015.) Disponible en: http://www.globalhealthaction.net/index.php/gha/article/view/5302
15. Lazer D, Kennedy R, King G, et al. The parable of Google flu: traps in big data analysis. Science [Internet]. 2014;343:1203–5 (Consultado el 10/05/2015.) Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/24626916
16. European Comission. Why do we need an EU data protection reform? [Internet]. 2011. p. 10-1. (Consultado el 10/05/2015.) Disponible en: http://ec.europa.eu/justice/data-protection/document/review2012/factsheets/1_en.pdf