Deriving information from external Big Databases and Big Data analytics: all that glitters is not gold

Miguel Angel Martinez-Garcia[1] and Anh Tuan Dinh-Xuan[2]

Affiliations: [1]Respiratory Dept, Hospital Universitario y Politécnico La Fe, Valencia, Spain. [2]Service de Physiologie, Paris Descartes University EA 2511, Hôpital Cochin, Assistance Publique Hôpitaux de Paris, Paris, France.

Correspondence: Miguel Angel Martinez-Garcia, Respiratory Dept, Hospital Universitario y Politécnico La Fe, Avenida Fernando Abril Martorell s/n 46026, Valencia, Spain. E-mail: mianmartinezgarcia@gmail.com

@ERSpublications
Big Data analytics are a further step on the road to the ideal model of personalised medicine
http://ow.ly/XWr1C

> *True genius resides in the capacity for evaluation of uncertain, hazardous, and conflicting information*
>
> Winston Churchill (1874–1965)

It seems more and more obvious that we are now living in the era of technology and information. Technological advances have allowed us to generate vast amounts of information of all kinds extremely quickly and, more importantly, store it so that it remains available for subsequent analysis. Accordingly, many large companies and institutions centre a large part of their strategy on the analysis of enormous databases on their customers. These provide information that determines future decisions and optimises resources.

For a long time, medical insurance companies have been handling huge quantities of information about millions of customers worldwide, stored in well-structured databases so that it can be analysed for business purposes. Such an enormous depository of information is also potentially useful for other purposes, such as the attainment of valuable scientific knowledge that would be very difficult to achieve by other means [1, 2]. This process does have a number of limitations, however, because these databases were not specifically created with scientific objectives in mind. The information they contain may not therefore be representative of the population under analysis. Alternatively, other conditioning factors may emerge, such as ethical problems derived from the use of personal data, a lack of key scientific information and insufficient quality control of the stored data [3–5].

In this issue of the *European Respiratory Journal*, Mokhlesi *et al.* [6] report the results of an interesting study that clearly shows some of the advantages and disadvantages of the use of large external databases that were not created for scientific purposes. This study used the MarketScan database (Truven Health Analytics, Ann Arbor, MI, USA), which holds information on 77.8 million working adults and retirees with employer-sponsored health insurance. After applying selection criteria, the authors managed to identify more than 1.7 million individuals with obstructive sleep apnoea (OSA), the largest number of OSA patients analysed in a scientific study to date, and a similar number of individuals with no evidence of OSA, who were chosen at random as a control group, adjusting for sex and age. The huge number of patients involved gives this study two significant strengths. On the one hand, the authors were able to

Received: Jan 09 2016 | Accepted: Jan 20 2016

Conflict of interest: None declared.

confirm, with great precision (a narrow confidence interval), some previously known associations between OSA and certain comorbidities, especially those of a cardiovascular and neurocognitive nature. On the other hand, and perhaps more importantly, the authors also achieved great statistical power to analyse particularly interesting subgroups of OSA patients that have scarcely been studied until now (for example, women and different age groups). This provides new information in this respect (figure 1 in their article is particularly enlightening).

Nevertheless, as the title of this editorial indicates, when it comes to using databases created for nonscientific purposes, and notwithstanding their enormous size, all that glitters is not gold, and this same study also sheds light on the limitations mentioned above, as acknowledged by the authors. Thus, as the databases were designed for medical insurance companies, most of the variables recorded are related to financial matters rather than scientific considerations. In other words, some variables that would be crucial to any scientific study of OSA patients were not collected and had to be approximately computed by indirect means. In this case, these variables would be the body mass index, race, demographic aspects, the severity of OSA, compliance with OSA treatment and the possible presence of patients with OSA in the control group (as not all the subjects in this group underwent a sleep study). Furthermore, the sample cannot be considered representative of the population being studied as it was taken from a database that only included working people. Finally, it must be stressed (although this is not the case here) that the use of large databases makes it more important than ever to distinguish between what is statistically significant and what is clinically relevant. It is well known that the more data that are available for comparison, the greater the probability of finding statistically significant associations, without them necessarily being clinically relevant [7]. As the conclusions of studies of this type could be considerably altered by some of these circumstances, they must be considered as nothing more than hypothesis-generating studies, *i.e.* studies capable of making new scientific discoveries but incapable of ratifying them, as this would require further studies specifically designed for this purpose.

Information is not knowledge; rather, it is the platform on which knowledge will be built. The amount of worldwide information that we have accumulated is now so vast and grows at such speed that power no longer resides with the holders of information but with whoever has the capacity to manage and process it as quickly and as efficiently as possible. This situation takes us to the heart of a phenomenon that has only hit the world of medicine in the last few years: Big Data analytics. There is still no generally accepted definition of Big Data but a good approximation would be "large volumes of high-velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information" [8]. According to this definition, the study by Mokhlesi *et al.* [6] cannot be fully considered a Big Data study because although the authors use a very "big" database, they do use the customary statistical tests to analyse their data. It has now been established that the average patient is capable of generating around 2 gigabytes of potentially analysable information, while an average hospital that in 2010 was handling an annual rate of 150–200 terabytes of information would deal with 600–700 terabytes in 2015 and would expect the rate of growth to be even faster in the future [9]. Everyday medical appointments, emergency visits, medical insurance, hospital admissions and the large amount of data provided by complementary tests, not forgetting the use of all the electronic connected devices (watches, mobile phones, sensors, *etc.*) capable of storing medical information, or the Internet of Things [10] (usually defined as the environment in which objects or people are provided with unique identifiers and the ability to transfer data over a network (normally the Internet) without requiring human-to-human or human-to-computer interaction), are all an inexhaustible source of present and future information on patients that is capable of generating new medical knowledge. It is widely recognised, for example, that Flu Trends (Google, Mountain View, CA, USA) succeeded in breaking the news of an influenza epidemic and tracking the distribution of the virus before any other communications media by analysing the millions of related searches undertaken on Google in the days beforehand (at least in its original model in 2008–2009; subsequent reports have asserted that Google Flu Trends' predictions had sometimes been very inaccurate) [11, 12]. Some authors have speculated that Big Data could have a significant number of uses in medicine, such as detection of diseases at earlier stages; management of specific individual and general health problems; identification of complications; analysis of disease progression, new associations and causal factors; generation of working hypotheses; more clinically relevant and cost-effective ways to diagnose and treat patients; predictive modelling; statistical tools to improve clinical trial designs; recognition of disease patterns; establishment of requirements and services for prediction and prevention; real-time analysis of the large volume of fast-moving data derived from in-hospital and in-home devices; identification of patients who are the greatest consumers of health resources; management of population health by detecting vulnerability variables; genomic analytics; evidence-based medicine; fraud analysis, *etc.* [8]. The McKinsey Global Institute estimates that Big Data analytics can enable a saving of more than 300 billion dollars per year in US healthcare, two thirds of which would come through reductions of approximately 8% in national healthcare expenditure [13]. One particular benefit in sleep medicine could result from the enormous quantity of data provided by night-time monitoring of polysomnographic variables or automatic continuous positive airway pressure devices, most of which we are unable to make use of in normal practice but which undoubtedly conceals highly valuable information, if only we could process it [14].

The time has now come to use all the technology for managing this prodigious mass of information to establish good logistical interconnections between all the parties that handle these vast databases [15]. However, we must not forget that "small data" often offer information that it is not contained in Big Data; therefore, perhaps it is time to focus in "all data revolution", using data from all traditional and new sources together [11]. This would guarantee high-quality information, unencumbered by noise; information that would incorporate the findings needed by database studies to come up with strong, valid conclusions and provide significant scientific knowledge. At the moment, technology offers us more and more pieces of the jigsaw but in an apparently chaotic fashion, and moreover, it seems to provide us with some pieces that cannot be put to use. Our task now is to put each piece in its rightful place and discard any that is merely noise so that this "jigsaw" works as a whole, with power and precision. Everything points to Big Data analytics being an important piece in this jigsaw, as well as a further step on the road towards the ideal and most expected model of personalised medicine.

References

1 Hsu YC. Analyzing Taiwan's National Health Insurance Research Database to explicate the allocation of health-care resources. *Adv Dig Med* 2015; 2: 41–42.

2 Warren JL, Klabunde CN, Schrag D, *et al.* Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care* 2002; 40: 3–18.

3 Nunan D, Di Domenico M. Market research and the ethics of big data. *Int J Mark Res* 2013; 55: 505–520.

4 Poynter R. 3 tips from Big Data from Nate Silver's 'The Signal and the Noise'. https://www.visioncritical.com/3-tips-big-data-nate-silver-s-signal-and-noise/ Date last updated: January 15, 2014. Date last accessed: January 2, 2016.

5 Tene O, Polonetsky J. Privacy in the Age of Big Data. A Time for Big Decisions. www.stanfordlawreview.org/online/privacy-paradox/big-data Date last updated: February 2, 2012. Date last accessed: January 2, 2016.

6 Mokhlesi B, Ham S, Gozal D. The effect of sex and age on the comorbidity burden of OSA: an observational analysis from a large nationwide US health claims database. *Eur Respir J* 2016; 47: 1162–1169.

7 Kieser M, Friede T, Gondan M. Assessment of statistical significance and clinical relevance. *Stat Med* 2013; 32: 1707–1719.

8 Raghupathi W, Raghupathi V. Big Data Analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014; 2: 3.

9 Pulido Cañabate, E. Big data: ¿solución o problema? [Big data: solution or problem?]. Madrid, Universidad Autónoma de Madrid, 2014.

10 Boulos MNK, Al-Shorbaji NM. On the Internet of Things, smart cities and the WHO Healthy Cities. *Int J Health Geogr* 2014; 13: 101.

11 Lazer D, Kennedy R, King G, *et al.* The parable of Google Flu: traps in Big Data analysis. *Science* 2014; 343: 1203–1205.

12 Ginsberg J, Mohebbi MH, Patel RS, *et al.* Detecting influenza epidemics using search engine query data. *Nature* 2009; 457: 1–5.

13 Manyika J, Chui M, Brown B, *et al.* Big data: the next frontier for innovation, competition, and productivity. www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation Date last updated: May 2011. Date last accessed: January 2, 2016.

14 Redline S, Dean D, Sanders H. Entering the era of "Big Data": getting our metrics right. *Sleep* 2013; 36: 465–469.

15 Dinh-Xuan AT. "Big data" and respiratory medicine. *Rev Mal Respir Actual* 2015; 7: 197–199.

DOI: 10.1183/13993003.00060-2016