ERS International Congress, Madrid, 2019: highlights from the Epidemiology and Environment Assembly

Mateusz Jankowski[1] and André F.S. Amaral[2]

Affiliations: [1]Dept of Epidemiology, School of Medicine in Katowice, Medical University of Silesia, Katowice, Poland. [2]National Heart and Lung Institute, Imperial College London, London, UK.

Correspondence: André F.S. Amaral, National Heart and Lung Institute, Imperial College London, Emmanuel Kaye Building, 1B Manresa Road, London SW3 6LR, UK. E-mail: a.amaral@imperial.ac.uk

ABSTRACT At the European Respiratory Society's International Congress of 2019, which was held in Madrid, Spain, there were several sessions with exciting poster and oral presentations within the fields of epidemiology and tobacco control. This article is the summary of two of these sessions. One was on the use of Big Data in epidemiology and the other, on the global burden of respiratory disease and tobacco.

@ERSpublications
Highlights from the Epidemiology and Environment Assembly at the #ERSCongress 2019 http://bit.ly/38p6jEZ

Cite this article as: Jankowski M, Amaral AFSERS International Congress, Madrid, 2019: highlights from the Epidemiology and Environment Assembly. *ERJ Open Res* 2020; 6: 00320-2019 [https://doi.org/10.1183/23120541.00320-2019].

Received: 19 Nov 2019 | Accepted: 11 Dec 2019

How do Big Data and classical epidemiology contribute to solving chronic respiratory ill health?

Big Data analytics in medicine refers to the high-speed analysis of large and complex datasets from thousands or millions of people, which allows us to identify clusters and correlations across datasets, as well as to develop predictive models using data mining techniques [1]. The aim of this session was to demonstrate how large datasets, including routinely collected electronic healthcare records, can be used to complement cohorts and more traditional epidemiology methods to answer a wide variety of questions about chronic respiratory diseases.

There are numerous observational studies on potential associations between exposures and asthma in children [2]. According to the first speaker of this session, Seif Shaheen, many of these associations are likely to be due to chance (random error), bias or confounding [3]. Shaheen talked about the need for more robust evidence to feed randomised controlled trials aiming to identify exposures that can improve lung function and prevent asthma early in life. From his point of view, Big Data, such as those linked to registers, should be used to improve the quality of epidemiological studies on asthma and lung function in children, as this would provide adequately powered samples to infer causality, using for example gene–environment interactions and Mendelian randomisation.

The second speaker, Catarina Almqvist Malmros, presented her approach to the study of early-life risk factors and consequences of asthma using large national population and health data registers. Almqvist Malmros mentioned several examples of high-quality registers from Sweden dedicated to specific diseases and how their good practice in data collection has proved very useful in many epidemiological studies. The linkage between registers and study cohort data together with the ability to generalise study findings to the national population are key strengths in the use of the Scandinavian Big Data for medical research. However, she also mentioned that the use of such Big Data presents challenges such as heavy computation needs, data integrity and security.

Sanja Stanojevic talked about the usefulness of large datasets in research on rare diseases, such as the Canadian, UK and US cystic fibrosis registers. According to Stanojevic, biobanks, Big Data phenotyping and machine learning algorithms can make a significant contribution to the level of knowledge on the diagnosis and treatment of rare diseases [4]. However, the innovation of these approaches will be a reality only when new questions are raised and data on alternative exposures are collected. One of the main limitations in using data from rare disease registers is oversampling of patients with more severe disease, and this is something for which classical epidemiology and Big Data have no solution.

Big Data can be also applied in environmental epidemiology. Roel Vermeulen presented a novel approach to the assessment of exposure to environmental pollutants, whereby the external and internal components of the exposome are characterised [5]. Vermeulen gave as an example of large-scale, real-time exposome assessment a study on which his team and Google collaborate to map ambient nitrogen dioxide levels in California (USA), the UK and Copenhagen (Denmark). He also gave examples of exposome assessment at the molecular level that uses metabolomics and epigenetic data. According to Vermeulen, the incorporation of exposomic data in epidemiological studies can improve the understanding of the relationship between risk factors and diseases, which can eventually lead to more effective prevention and control.

Finally, Deborah Jarvis talked about the potential of implementing Big Data analysis methods in the context of biomedical and health research, with particular emphasis on cohort studies. Jarvis presented two examples of Big Data analyses; one was Google Flu Trends and its pitfalls [6], and the other was a study where data on outdoor NO_2 and low birthweight from the UK birth register and the Millenium cohort study were jointly analysed [7]. She also mentioned the usefulness of linking electronic health records to cohorts and large biobanks.

Big Data and machine learning algorithms integrated with classical epidemiology and cohort data may make a significant contribution to the research in the field of respiratory diseases. However, to use the potential of large datasets fully, several actions and guidelines in the fields of data availability and linkage to registers, data quality and security, as well as dedicated regulatory policies are needed.

Global burden of respiratory disease and tobacco: an update

The Global Burden of Disease (GBD) programme is a good example of how Big Data can be used to estimate the current and future impact of diseases such as COPD and asthma [8] at a global scale. This session provided a great overview of recent findings of the GBD programme on chronic respiratory disease, with a special focus on tobacco smoking, as well as a critical view of the method used by this programme to estimate the burden of diseases.

Vinay Gupta talked about the current state of work undertaken by the GBD programme on chronic respiratory disease and smoking, and how this will progress in the future. Throughout his presentation, Gupta provided us with estimates of the burden of chronic respiratory disease for several regions of the world. He reminded us that the highest prevalence of chronic respiratory disease is seen in high-income countries, with COPD and asthma being the leading causes of death due to respiratory disease. However, with a greater number of chronic respiratory disease-attributable years of life lost in South Asia, it seems that in this low-income region, people dying of chronic respiratory disease are younger than in other regions [8]. Although COPD and asthma are still the most prevalent diseases across all regions, Gupta urged we should pay attention to interstitial lung disease and sarcoidosis, which are becoming more prevalent, especially among men. As tobacco smoking is the leading risk factor for chronic respiratory disease, the GBD programme has put a lot of effort into getting the best estimates of smoking prevalence and risk of disease associated with this behaviour. Gupta informed us that the method to estimate smoking risk in GBD is changing to include occasional and former smokers, and cumulative exposure. They aim to incorporate also data on electronic cigarette use, maternal smoking, exposure to second-hand smoke and policy measures applied across countries that could impact on the estimates over time. In addition, they wish to interpret effects of past policies and forecast the future if one of several scenarios (*e.g.* tax increases, moving from traditional cigarettes to e-cigarettes and plain packaging) were implemented.

Luisa Sorio Flor briefly described the method used by the GBD programme to estimate the burden of disease due to tobacco smoking. The GBD programme estimates smoking prevalence and individual-level exposure distributions, the dose–response risk curve for current smokers and risk reduction curve for former smokers, as well as the theoretical minimal risk exposure level (*i.e.* if the entire population had never smoked). Smoking prevalence data, which are mainly based on high-income and some middle-income countries (Africa and Asia are extremely underrepresented), are obtained from the literature and collaborating studies, and then extrapolated to all countries of the world and all age groups. The same approach is used to estimate cumulative exposure (*i.e.* pack-years). After identifying possible smoking-outcome pairs for 36 diseases, the GBD programme estimated dose–response curve relative risks (RRs), and based on these RRs and prevalence of smoking, they estimated the population attributable fraction due to tobacco smoking. According to Gupta and Sorio Flor, the number of deaths attributable to tobacco smoking in 2017 was 7.1 million, or 8.1 million if second-hand smoke is considered, making this the second-leading risk factor for early all-cause mortality in 2017 (12% of all deaths). In 2020, the GBD programme will implement a scoring system to the evidence on risk–disease pair and relative risk estimation.

The last speaker of this session was Peter Burney, who provided a critical view of the GBD method. Burney pointed out that the data used by the GBD programme come from several studies not necessarily using the same definition of outcome, with some based on diagnoses (*e.g.* doctor-diagnosed asthma or COPD) and others based on objectively measured conditions (*e.g.* lung function). This could partly explain the high heterogeneity across studies in several of the systematic reviews and meta-analyses used by the GBD programme in their estimates. As an example of this issue, Burney showed how the estimate of burden of chronic respiratory disease associated with the use of biomass is likely to be inflated, as the estimate of RRs used by the GBD programme are based on biased and highly heterogeneous data. Burney also questioned whether the high mortality reported by the GBD for low- and middle-income countries (LMICs) is really due to chronic airflow obstruction. He said that mortality is most likely due to poor lung development, although data on this issue are still lacking. Other key limitations that Burney mentioned were the lack of death certificates in poor countries and the lack of precision in the air pollution literature, which does not consider the different effects of the several components of air pollution (*e.g.* biological fraction of particulate matter).

Despite the flaws of the GBD programme in estimating the burden of disease, it is an important public health initiative. The GBD estimates would likely improve if more primary data were collected in LMICs and if only high-quality data were included in their models.

Concluding remarks

These two sessions showed where epidemiology of respiratory diseases is heading, and where the strengths and limitations of using Big Data lie, particularly at the national or global scale. We hope that this summary presented through the Epidemiology and Environment Assembly of the ERS generates curiosity in readers, especially early career epidemiologists, to follow-up on these topics.

Conflict of interest: M. Jankowski has nothing to disclose. A.F.S. Amaral works in the same group as two of the speakers (D. Jarvis and P. Burney) whose talks are commented on.

References

1 Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform* 2015; 19: 1209–1215.

2 Castro-Rodriguez JA, Forno E, Rodriguez-Martinez CE, *et al.* Risk and protective factors for childhood asthma: what is the evidence? *J Allergy Clin Immunol Pract* 2016; 4: 1111–1122.

3 Shaheen S. Elucidating the causes of asthma: how can we do better? *Lancet Respir Med* 2019; 7: e25.

4 Hallowell N, Parker M, Nellaker C. Big data phenotyping in rare diseases: some ethical issues. *Genet Med* 2019; 21: 272–274.

5 Vineis P, Chadeau-Hyam M, Gmuender H, *et al.* The exposome in practice: design of the EXPOsOMICS project. *Int J Hyg Environ Health* 2017; 220: 2 Pt A, 142–151.

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

7 Jackson CH, Best NG, Richardson S. Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics* 2009; 10: 335–351.

8 GBD 2015 Chronic Respiratory Disease Collaborators. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Respir Med* 2017; 5: 691–706.