Respiratory Development: Thinking Outside the Box - Big Data for Respiratory Medicines ?

David Mannino[1,2]

[1]GlaxoSmithKline, 5 Crescent Drive, Philadelphia, PA, 19112, USA
[2]University of Kentucky, 111 Washington Avenue, Lexington, KY, USA

Summary

The concept of big data, which is typically characterized by volume, variety, veracity, and velocity has the capacity to transform the scientific approach to respiratory disease. This transformation ranges from a better understanding of the public health or population health approach to disease, to a drug safety and drug development, to alternative approaches to monitoring outbreaks and epidemics. Medical big data can generally be thought of as being in one of three classes: large numbers (often millions) of people with small numbers of parameters (such as mortality data or other administrative data); smaller numbers of people with large amounts of data (such as micro array or genetic data); and, more recently large numbers of people with large amounts of data. Each of these classes present their own challenges and opportunities. For example, administrative data is often complicated by issues such as missing or incorrect data and potential biases, such as residual confounding or reverse causality. In addition, big data may be useful for hypothesis generation but is not rally able to test causality. Big data approaches have been used in respiratory in several applications: identification of asthma mortality patterns across different countries over many years (which pointed to certain classes of medications as being responsible); determining the relation between area of residence and respiratory mortality (and those changes over time); and identification of the relation between air pollution exposure and mortality. Future applications may help to define better targets for respiratory therapy development.

Key Message

Big data provides both promise and challenges in better understanding trends and patterns of respiratory disease, along with drug development and safety. The promise of big data includes identifying patterns of disease, finding potential drug targets, and improving the efficiency of new discoveries. The challenges of big data include missing or incorrect data and the inability to test causality with it.

Introduction

This paper reviews the role of big data in the understanding of respiratory disease and its potential role in the drug development and drug safety process. The components of this paper include:

1. Defining medical big data
2. The promise of big data
    a. Understanding asthma mortality
        i. International comparisons over time
        ii. Small area changes in asthma deaths over time
    b. Treatment of COPD and Asthma in a real world setting
        i. Salford lung study
    c. Identifying potential drug targets or genetic variants
        i. Asthma example
3. The challenges of big data
    a. Missing and dirty data
    b. Bigger is not always better
        i. Cautionary tale of influenza tracking

Defining Medical Big Data

The concept of “medical big data” has emerged in recent years as means of transforming medicine.[1] This data, by itself, is not transformative, but its subsequent analysis, interpretation, and the actions that may follow. The factors that make data “big” include the 4 V’s - volume, variety, velocity, and veracity [2] is important to consider in the application to healthcare. This provides the potential to represent what happens, in a nearly unbiased way, what happens in the real world. Medical big data can arise from a number of different sources, including public health databases (birth and death certificates), administrative claims data, clinical registries, biometric data, electronic medical records, patient reported data, imaging data, clinical trials, and prospective cohorts, to name several sources.[3]

Medical big data can be broadly classified in to one of three forms, based on sample numbers and parameter numbers: those with very large samples and a small number of parameters, those with smaller samples and a very larger number of parameters, and, more recently, those with large samples and larger parameters. The first example has been used in public health applications for many years to analyze mortality and administrative data, and classical statistical techniques can typically be applied in the analyses. The second classification is newer and can be seen in applications where hundreds or thousands of parameters are evaluated in individuals (such as genetic data from micro-arrays). Classical statistical testing is typically not useful in the interpretation of this type of data. The final type of big data, comprising large amounts of data on large numbers of people, present some of the issues seen in both the first and second types as described above.

The Promise of Big Data

Some of the potential value of medical big data has been demonstrated in the delivery of personalized medicine, the use of clinical decision support systems, tailoring diagnostic and treatment decisions, data driven population health analyses using large databases, and fraud detection and prevention.[4] Rumsfeld summarized eight areas of application of big data analytics in the improvement of healthcare as predictive modelling for risk and resource use, population management, drug and medical device safety surveillance, disease and treatment heterogeneity, precision medicine and clinical decision support, quality of care and performance measurement, public health, and research applications.[5] Some of these applications have proven useful in the study of respiratory disease.

Asthma-related mortality provides two examples. A recent study looked at over 50 years of asthma deaths from 46 countries. [6] While there has been a downward trend over the this time period, there is also considerable differences between countries. Of particular interest is the historical peaks in asthma mortality observed in New Zealand in the 1960s and again in the 1970s and 1980s. This was thought to be due to the overuse of the beta agonists isoprenaline in the former period and fenoterol in the later period. A separate study examined asthma mortality over a 35 year period in the US counties.[7] This study found an overall decrease in asthma deaths over this period of time. The exception to this decrease occurred in a swath of counties in the southern part of the US that represents a population where poor blacks are overrepresented. Thus, this study points toward social and economic factors being important in asthma mortality.[8]

A different kind of example of the use of big data in respiratory medicine is the Salford Lung Study, which was designed as a pragmatic, randomized real-world effectiveness study. [9] This unique study recruited patients from clinical practices, but followed them using electronic medical records and administrative data (rather than the classical follow-up using study monitors). The COPD trial found a reduced rate of exacerbations in the intervention group (by 8.4%).[10] The asthma study found a higher rate of response in the intervention group (71% vs. 56%).[11] Both of these trials used big data to examine effectiveness of therapy as opposed to efficacy, which is what standard randomized controlled trials can test.[12]

Another example of how big data can be used to inform response to therapeutic agents is the application of genetic analysis to a cohort of children with asthma to determine what factors might explain bronchodilator response.[13] In this study, five different loci associated with bronchodilator responsiveness were identified, but these showed substantial differences when looking at ethnic subgroups (Puerto Ricans, African Americans, and Mexicans).

The Challenges of Big Data

Medical big data also can present significant challenges.[3] In the real world, data may be missing or incorrect. In classical statistical analyses subjects with missing data are often excluded from the analysis- although this assumes subjects with missing data do not differ from those without missing data. In the real world of big data, this is assumption is probably not true (i.e. those with complete data and follow-up may not be similar to those with missing data and incomplete follow-up). Thus, different analytic techniques need to be used to account for this difference.

Another challenge in big data analysis is the "curse of dimensionality", which a problem seen when there are too many attributes relative to the number of observational units. This can manifest as multicolinearity, where two or more variables in a model are not independent. Although there are techniques designed to deal with this problem, they can result in the loss of important information.[3]

Another challenge in the world of big data is avoiding the trap that " bigger is better" and that these newer approaches provide more valuable information compared to traditional approaches. An example of this phenomenon can be seen in the story of Google Flu Trends, aninfluenza tracking system that was thought to be a better way for tracking flu outbreaks compared to the traditional surveillance reports that the US Centers for Disease Control and Prevention (CDC) has been using for many years. In late 2012 and early 2013, the Google Flu Trends estimates of flu prevalence ended up being more than double what CDC had estimated.[14] While this was thought to be related to issues in the algorithm used to determine flu, an important factor was thought to be a function of the "media-stoked panic" related to influenza during the 2012-2013 flu season.[14] In this example, the traditional approach that CDC has used for years proved more accurate, although taking advantage of other approaches, including Google Flu Trends, may provide some additional unique and helpful information.

Finally, in most application, medical big data is useful in hypothesis generation but is not able to demonstrate causality

Summary

Big data provides both promise and challenges in better understanding trends and patterns of respiratory disease, along with drug development and safety. The promise of big data includes identifying patterns of disease, finding potential drug targets, and improving the efficiency of new discoveries. The challenges of big data include missing or incorrect data and the inability to test causality with it.

References

1. Obermeyer Z, Emanuel EJ. *Predicting the Future - Big Data, Machine Learning, and Clinical Medicine*. N Engl J Med. 2016;375:pp 1216-9.

2. Bellazzi R. *Big data and biomedical informatics: a challenging opportunity*. Yearb Med Inform. 2014;9:pp 8-13.

3. Lee CH, Yoon HJ. *Medical big data: promise and challenges*. Kidney Res Clin Pract. 2017;36:pp 3-11.

4. Roski J, Bo-Linn GW, Andrews TA. *Creating value in health care through big data: opportunities and policy implications*. Health Aff (Millwood). 2014;33:pp 1115-22.

5. Rumsfeld JS, Joynt KE, Maddox TM. *Big data analytics to improve cardiovascular care: promise and challenges*. Nat Rev Cardiol. 2016;13:pp 350-9.

6. Ebmeier S, Thayabaran D, Braithwaite I, Benamara C, Weatherall M, Beasley R. *Trends in international asthma mortality: analysis of data from the WHO Mortality Database from 46 countries (1993-2012)*. Lancet. 2017;390:pp 935-45.

7. Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, Morozoff C, Shirude S, Naghavi M, Mokdad AH, Murray CJL. *Trends and Patterns of Differences in Chronic Respiratory Disease Mortality Among US Counties, 1980-2014*. JAMA. 2017;318:pp 1136-49.

8. Mannino DM, Sanderson WT. *Using Big Data to Reveal Chronic Respiratory Disease Mortality Patterns and Identify Potential Public Health Interventions*. JAMA. 2017;318:pp 1112-4.

9. Bakerly ND, Woodcock A, New JP, Gibson JM, Wu W, Leather D, Vestbo J. *The Salford Lung Study protocol: a pragmatic, randomised phase III real-world effectiveness trial in chronic obstructive pulmonary disease*. Respir Res. 2015;16:pp 101.

10. Vestbo J, Leather D, Diar Bakerly N, New J, Gibson JM, McCorkindale S, Collier S, Crawford J, Frith L, Harvey C, Svedsater H, Woodcock A, Salford Lung Study I. *Effectiveness of Fluticasone Furoate-Vilanterol for COPD in Clinical Practice*. N Engl J Med. 2016;375:pp 1253-60.

11. Woodcock A, Vestbo J, Bakerly ND, New J, Gibson JM, McCorkindale S, Jones R, Collier S, Lay-Flurrie J, Frith L, Jacques L, Fletcher JL, Harvey C, Svedsater H, Leather D, Salford Lung Study I. *Effectiveness of fluticasone furoate plus vilanterol on asthma control in clinical practice: an open-label, parallel group, randomised controlled trial*. Lancet. 2017;390:pp 2247-55.

12. Ryan D, Blakey J, Chisholm A, Price D, Thomas M, Stallberg B, Lisspers K, Kocks JWH, Respiratory Effectiveness G. *Use of electronic medical records and biomarkers to manage risk and resource efficiencies*. Eur Clin Respir J. 2017;4:pp 1293386.

13. Mak ACY, White MJ, Eckalbar WL, Szpiech ZA, Oh SS, Pino-Yanes M, Hu D, Goddard P, Huntsman S, Galanter J, Wu AC, Himes BE, Germer S, Vogel JM, Bunting KL, Eng C, Salazar S, Keys KL, Liberto J, Nuckton TJ, Nguyen TA, Torgerson DG, Kwok PY, Levin AM, Celedon JC, Forno E, Hakonarson H, Sleiman PM, Dahlin A, Tantisira KG, Weiss ST, Serebrisky D, Brigino-Buenaventura E, Farber HJ, Meade K, Lenoir MA, Avila PC, Sen S, Thyne SM, Rodriguez-Cintron W, Winkler CA, Moreno-Estrada A, Sandoval K, Rodriguez-Santana JR, Kumar R, Williams LK, Ahituv N, Ziv E, Seibold MA, Darnell RB, Zaitlen N, Hernandez RD, Burchard EG, Consortium NT-OfPM. *Whole-Genome Sequencing of Pharmacogenetic Drug Response in Racially Diverse Children with Asthma*. Am J Respir Crit Care Med. 2018;197:pp 1552-64.

14. Lazer D, Kennedy R, King G, Vespignani A. *Big data. The parable of Google Flu: traps in big data analysis*. Science. 2014;343:pp 1203-5.