Diagnostic Microbiology and Infectious Disease 94 (2019) 209–210

Contents lists available at ScienceDirect

Diagnostic Microbiology and Infectious Disease

journal homepage: www.elsevier.com/locate/diagmicrobio

Letter to the Editor

Pitfalls in big data analysis: next-generation technologies, last-generation data

Traditionally, scientific discoveries were usually achieved by the classical sequence of "generation of hypothesis–collection of high-quality specific data for testing the hypothesis–data analysis–conclusion" approach in *The Scientific Method*. In recent years, the exponential increase in computational power as predicted by Moore's law and the accumulation of a large amount of data throughout the years have facilitated the emergence of a robust discipline called big data analysis. The big data used for analysis are often pooled retrospective data collected by multiple parties, which may involve hundreds and thousands of individuals or parties with varying qualities and training. These data are stored in databases that are publicly accessible or owned by an organization but can be readily retrieved by individuals of the organization. Through the examination of these large data sets with improved statistical power, big data analysis is able to uncover new patterns and associations, and hence draw new conclusions, that traditional analysis of smaller data sets generated by individual research groups cannot achieve. However, such discovery of new associations relies heavily on the availability of high-quality data sets, which may not be the case for data stored in large databases, as the data qualities are often not well controlled during the data collection process by the individual parties that contribute to the content of the databases. In most studies, the occasional suboptimal data qualities can be diluted as a result of the large quantity of data, but this may not be the case in some others. On the other hand, prospectively collected big data from multicenters are usually of higher quality, as the protocols of data collection are more standardized (Michaud et al., 2013).

In clinical microbiology and infectious disease, one of the most frequently performed big data analyses is to examine the association between the presence of specific microbes and a particular disease of interest. Sometimes, the association may imply a causal relationship between the microbe and the disease. In other occasions, the microbe can be a marker for a particular disease such that detection of the microbe from a patient may warrant further investigations for and guide early diagnosis of the disease. The emergence of these unprecedented discoveries has resulted in an explosion of scientific publications using big data, some of which reported in leading journals (Dellon et al., 2011; Kwong et al., 2018; Orlovska et al., 2017). In one study, using the US pathology database, Dellon et al. discovered that *Helicobacter pylori* infection was inversely associated with esophageal eosinophilia, which might offer clues on the epidemiology and pathogenesis of esophageal eosinophilia (Dellon et al., 2011). In another study, using the nationwide Danish database, Orlovska et al. found that streptococcal throat infections were associated with increased risks of mental disorders, particularly obsessive–compulsive and tic disorders, which may support the diagnostic concept of pediatric acute-onset neuropsychiatric syndrome (Orlovska et al., 2017). These studies, with reliable and accurate microbial data available for analysis, well illustrate the power of big data for searching and establishing new disease associations with microbes as risk factors. On the other hand, the conclusions made can be in doubt if big data with suboptimal quality were used for analysis. For example, in a recent study which investigated the association between bacteremia due to specific intestinal microbes and colorectal cancer (Kwong et al., 2018), the data (2006–2015) were retrieved from a centralized computerized database of electronic medical records managed by the Hong Kong Hospital Authority. The microbiological data obtained included specific bacterial genus and/or species reported in the blood cultures of patients, which were input to the database by technicians of clinical microbiology laboratories in public hospitals in Hong Kong as part of their routine microbiology service. The study found that colorectal cancer was associated with certain genus and species of bacteria, such as *Bacteroides fragilis*, *Streptococcus gallolyticus*, *Fusobacterium nucleatum*, *Peptostreptococcus* species, *Clostridium septicum*, *Clostridium perfringens*, and *Gemella morbillorum*. However, the conclusions made has not taken into account whether the reported blood culture isolates represent genuine bacteremia or contaminants and, more importantly, the limited accuracy of bacterial identification in clinical laboratory settings during the study period, especially for some groups of bacteria.

Before the wide use of matrix-assisted laser desorption ionization time-of-flight mass spectrometry in the last few years, identification of blood culture isolates in most clinical microbiology laboratories, including those in Hong Kong, was performed by conventional biochemical methods coupled with commercial bacterial identification platforms, such as the VITEK system (BioMerieux Vitek, Hazelwood, MO), the API system (BioMerieux Vitek), and the ATB Expression system (BioMerieux Vitek). 16S rRNA gene sequencing, the gold standard for bacterial identification and classification, was seldom performed in clinical microbiology laboratories (Woo et al., 2008). However, it is well known that the performance of phenotypic methods is often suboptimal for certain bacterial genera and species. From our previous works on 16S rRNA gene sequencing of archived bacterial isolates in our clinical microbiology laboratory, we found that the identities of many difficult-to-identify bacterial groups reported to the clinicians were often incorrect. For example, some *Streptococcus* species and *Streptococcus*-like organisms (e.g., *Gemella*, *Abiotrophia*, *Granulicatella*, *Helcococcus*) are often incorrectly identified by phenotypic methods and reported (Woo et al., 2003a, 2003b, 2005b). Similarly, anaerobic bacteria such as *Clostridium* and non–spore-forming anaerobic Gram-positive and Gram-negative bacilli are also frequently misidentified (Lau et al., 2004, 2006a, 2006b; Woo et al., 2002, 2005a, 2007). Therefore, reported bacterial names that were retrieved from the database may not be their true identities. Such microbiological data inaccuracies might seriously affect the conclusions drawn and should be taken into account in similar big data studies.

To ensure the accuracies of results obtained, big data analysts must be discreet in choosing high-quality data for their analysis especially if the data are retrospectively collected with the use of next-generation technologies for analysis of last-generation data. They should also address the limitations of the data quality while drawing conclusions so that readers are aware of such factors and their potential impact on the conclusions made, which may better be regarded as preliminary hypothesis requiring further studies to test its validity.

Acknowledgment

We thank Dr. Chi-Ching Tsang for his critical comments on the manuscript.

Susanna K.P. Lau*,
Patrick C.Y. Woo*,
Department of Microbiology, The University of Hong Kong, Hong Kong, China
State Key Laboratory of Emerging Infectious Diseases, The University of Hong Kong, Hong Kong, China
Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The University of Hong Kong, Hong Kong, China
*Corresponding authors. Tel.: +852-22554892; fax: +852-28551241.
E-mail addresses: skplau@hku.hk (S.K.P. Lau),
pcywoo@hku.hk (P.C.Y. Woo),

https://doi.org/10.1016/j.diagmicrobio.2018.12.006

References

Dellon ES, Peery AF, Shaheen NJ, Morgan DR, Hurrell JM, Lash RH, et al. Inverse association of esophageal eosinophilia with *Helicobacter pylori* based on analysis of a US pathology database. Gastroenterology 2011;141(5):1586–92.

Kwong TNY, Wang X, Nakatsu G, Chow TC, Tipoe T, Dai RZW, et al. Association between bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. Gastroenterology 2018;155(2):383–390.e8.

Lau SKP, Ng KHL, Woo PCY, Yip K-T, Fung AMY, Woo GKS, et al. Usefulness of the MicroSeq 500 16S rDNA bacterial identification system for identification of anaerobic gram positive bacilli isolated from blood cultures. J Clin Pathol 2006;59 (2):219–22.

Lau SKP, Woo PCY, Fung AMY, Chan K-M, Woo GKS, Yuen K-Y. Anaerobic, non-sporulating, gram-positive bacilli bacteraemia characterized by 16S rRNA gene sequencing. J Med Microbiol 2004;53(12):1247–53.

Lau SKP, Woo PCY, Li NKH, Teng JLL, Leung K-W, Ng KHL, et al. *Globicatella* bacteraemia identified by 16S ribosomal RNA gene sequencing. J Clin Pathol 2006;59(3):303–7.

Michaud DS, Izard J, Wilhelm-Benartzi CS, You D-H, Grote VA, Tjønneland A, et al. Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large European prospective cohort study. Gut 2013;62(12):1764–70.

Orlovska S, Vestergaard C, Bech B, Nordentoft M, Vestergaard M, Benros M. Association of streptococcal throat infection with mental disorders: testing key aspects of the pandas hypothesis in a nationwide study. JAMA Psychiat 2017;74(7):740–6.

Woo PCY, Fung AMY, Lau SKP, Chan BYL, Chiu SK, Teng JLL, et al. *Granulicatella adiacens* and *Abiotrophia defectiva* bacteraemia characterized by 16S rRNA gene sequencing. J Med Microbiol 2003;52(2):137–40.

Woo PCY, Fung AMY, Lau SKP, Yuen K-Y. Identification by 16S rRNA gene sequencing of *Lactobacillus salivarius* bacteremic cholecystitis. J Clin Microbiol 2002;40(1): 265–7.

Woo PCY, Lau SKP, Chan K-M, Fung AMY, Tang BSF, Yuen K-Y. *Clostridium* bacteraemia characterised by 16S ribosomal RNA gene sequencing. J Clin Pathol 2005;58(3):301–7.

Woo PCY, Lau SKP, Fung AMY, Chiu SK, Yung RWH, Yuen K-Y. *Gemella* bacteraemia characterised by 16S ribosomal RNA gene sequencing. J Clin Pathol 2003;56(9):690–3.

Woo PCY, Lau SKP, Lin AWC, Curreem SOT, Fung AMY, Yuen K-Y. Surgical site abscess caused by *Lactobacillus fermentum* identified by 16S ribosomal RNA gene sequencing. Diagn Microbiol Infect Dis 2007;58(2):251–4.

Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen K-Y. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect 2008;14(10):908–34.

Woo PCY, Tse H, Wong SSY, Tse CWS, Fung AMY, Tam DMW, et al. Life-threatening invasive *Helcococcus kunzii* infections in intravenous-drug users and *ermA*-mediated erythromycin resistance. J Clin Microbiol 2005;43(12):6205–8.