Stieb et al. BMC Public Health (2017) 17:372
DOI 10.1186/s12889-017-4286-8

BMC Public Health

EDITORIAL Open Access

Promise and pitfalls in the application of big data to occupational and environmental health

David M. Stieb[1,2*], Cécile R. Boot[3] and Michelle C. Turner[4,5,6,7]

Is "big data" merely a catchphrase, or does the approach hold real promise in informing occupational and environmental health? Can challenges related to messy and unrepresentative data and spurious findings be overcome?

Promise

The potential power of big data to inform public health decision-making has been widely recognized [1, 2]. However, there is a paucity of published primary research employing these methods in this journal and elsewhere [3, 4]. The *American Journal of Public Health* encouraged new research in this area and recently appointed an inaugural associate editor for digital health [3].

Big data are typically defined in relation to the "three Vs", volume, velocity and variety (and more recently, variability, veracity and value) [5]. Other defining characteristics include the emergence of new data sources and providers such as social media, mobile applications and wearable technology such as fitness trackers (the "quantified self" [6]), the need for new analytical methods such as machine learning, non-traditional multi-disciplinary partnerships and real-time analysis and forecasting [7].

Along similar lines, sharing of clinical trial and other study data has also been advocated as a means of broadening access to and more fully exploiting the collective power of data. In addition to increasing statistical power, which could potentially facilitate detecting small signals earlier, which may be particularly important in environmental health, advantages of pooling data include enhanced ability to examine heterogeneity between diverse populations, and consideration of novel hypotheses not tested by the original investigators [8]. Data sharing initiatives must overcome barriers including providing protections for original investigators, particularly those in low-resource countries [9], and issues related to data ownership, privacy and security [8]. The Healthy Birth, Growth, and Development–Knowledge Integration initiative is an example of a data sharing initiative which has navigated many of these issues [8]. A need has also been identified to address barriers to the international sharing of routinely collected public health data, including technical, motivational, economic, political, legal and ethical factors [10].

Exposure analysis is the keystone of occupational and environmental health. As a result, the concept of big data in this context is linked closely to that of the exposome, the totality of human environmental, occupational and other exposures from conception to death [11]. These exposures interact with other determinants of internal dose and health effects characterized by their own data-rich "omes" – the genome, metabolome, lipidome, transcriptome and proteome, among others, analysis of all of which requires novel data analysis methods [11–14]. The exposome may be characterized using a vast array of methods including measurement of both exogenous and endogenous biomarkers in biological specimens, direct environmental monitoring using dedicated sensors, and indirect sources such as operational data from metering and energy use, and facilities management data [12, 15–17].

Pitfalls

As a counterpoint to the potential of big data, one of the primary concerns is the potential for spurious findings, (described at their worst as "fanciful rubbish" or "big error") that can be generated by employing "much bigger and messier data" [2, 7]. Related to these limitations of big data are epistemological issues around the approach to how they are analyzed and how knowledge is generated. Some have gone so far as to argue that big data analytics allow the data to "speak for themselves," free of a priori hypotheses, and by extension of investigator bias, but others have countered that whether desirable or not, this is unattainable since all

* Correspondence: dave.stieb@hc-sc.gc.ca
[1]Population Studies Division, Environmental Health Science and Research Bureau, Health Canada, 420-757 West Hastings St. - Federal Tower, Vancouver, BC V6C 1A1, Canada
[2]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada
Full list of author information is available at the end of the article

data are in fact framed by the methods and constructs under which they are collected [2, 18]. A hybrid approach has been advanced where big data analysis, machine learning or "knowledge discovery" is guided by theory and practical experience, including a more selective approach to choosing appropriate data sources and analysis methods, as well as ultimately testing hypotheses generated from initial analyses [2, 18]. An additional concern is that to the extent that big data relies on consumer "data trails," mobile devices, wearable technology or electronic medical records, they may exclude those with limited footprints owing to barriers related to age, race, socioeconomic status, access to care or health literacy [5]. This has the potential to amplify environmental injustice concerns to the extent that it further disadvantages populations who already experience a disproportionate health burden related to environmental exposures [19].

Application to occupational and environmental health

Notwithstanding these important caveats, the potential for big data to inform public health and occupational and environmental health more specifically has been recognized by several funding agencies. The National Institute of Environmental Health Sciences is part of a National Institutes of Health-wide data science initiative, "Big data to knowledge" (BD2K), which aims to facilitate wide use of data, develop methods, software and tools, build capacity through training, and support data infrastructure [20]. The European Commission recently issued a call for proposals pertaining to "Big data supporting Public Health Policies," focusing on "how to better acquire, manage, share, model, process and exploit" big data for public health purposes, highlighting the opportunities they may provide to identify interactions between environmental, genetic and behavioral determinants of health [21]. Funded initiatives include the European Exposome Cluster [22], US Health and Exposome Research Center: Understanding Lifetime Exposures (HERCULES) [23], and the CANadian Urban Environmental (CANUE) Health Research Consortium [24].

Research in both occupational and environmental health has made widespread use of large datasets for many years. It is instructive to consider how it has been transformed by increasing application of big data and data sharing. In the environmental health realm, there is a long history in air pollution epidemiology of combining routinely available administrative health or vital statistics data, with environmental monitoring data, particularly to examine effects of short term variability in exposure using time-series or case-crossover analysis [25]. This approach was subsequently applied to examining the effects of long term exposure by linking an existing cohort, the American Cancer Society cohort [26], to routinely available environmental data, in order to relatively inexpensively replicate findings from a dedicated cohort study, the Six Cities Study [27]. This approach has now been applied to many other cohorts, and further by creating synthetic cohorts by linking census or tax data to vital statistics data and incorporating spatially comprehensive exposure data combining ground based monitoring, satellite observations, chemical/meteorological models and land use patterns [28, 29]. There are also examples of exploiting clinical trial data to examine associations with air pollution, unrelated to the original study hypothesis, e.g. linking clinical data on carotid intima media thickness as a measure of development of atherosclerosis, to air pollution exposure [30]. While social media as a source of big data have been dismissed as "frivolous," in addition to being used to track communicable disease for surveillance purposes, there are examples of application to chronic disease and environmental health such as development of predictive models of asthma using Twitter, Google searches and air monitoring data [31]. Asthma exacerbations are well documented in relation to air pollution exposure, and asthma also lends itself to "self-quantification" in relation to tracking of lung function and symptoms. Licksai et al. [32] developed a mobile application which combines these features of asthma with air quality forecasts and advice.

Similarly, in occupational health, workplace injury and illness data from physician reporting, employer records and workers compensation claims have been a longstanding resource for research and surveillance. Recently, the US Occupational Safety and Health Administration strengthened reporting requirements and improved public access to these data, motivated partly by increasing the utility of the data for research [33]. In Europe, investigators employed 20 physician reporting and compensation claim datasets from 10 countries to examine trends in occupational disease incidence, accounting for the diversity of data collection methods employed in each country, and demonstrated the potential of data sharing in this area [34]. A key aim of exploiting these data is to improve the capacity to predict and prevent injury and disease in the workplace [35]. Evaluating longer term sequelae of workplace disease and injury requires different types of data. Scandinavia has a long tradition of linking cohort studies to register data to gain insight into predictors of sick leave and work disability [36]. The social security system is a determining factor for the content of registers and there may be important differences between countries. While sick leave benefits are taken over by the social security system in Scandinavia relatively early in the process, in contrast in the Netherlands, the employer is responsible for payment of salary during the first two years of sick leave. As a result, there is no national registration of sick leave, which is a disincentive for employers for valid company registration, reducing its validity as a measure. Nonetheless, first attempts are being made in the Netherlands to

link occupational health cohort data to national registers that are a reliable source for measures related to source of income [37]. Social security data have also been widely used to examine work disability benefits and transitions from work to retirement.

Conclusions

Big data and data sharing have the potential to inform occupational and environmental health by exploiting innovations related to non-traditional data sources or providers and novel partnerships. Promising applications include real time analysis and forecasting, and innovative analyses of clinical trial or observational data originally collected for other purposes. However, in order to support these innovations, advances are also required in data curation, protection of privacy and security, as well as data analysis methods. Challenges related to messy and unrepresentative data and spurious findings, as well as epistemological issues and equity considerations must also be addressed.

Abbreviations

BD2K: Big Data to Knowledge; CANUE: Canadian Urban Environmental Health Research Consortium); HERCULES: Health and Exposome Research Center: Understanding Lifetime Exposures

Acknowledgements

Not applicable.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Authors' contributions

DMS, CRB and MCT were involved in drafting the manuscript or revising it critically for important intellectual content, participated sufficiently in the work to take public responsibility for appropriate portions of the content, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

Author details

[1]Population Studies Division, Environmental Health Science and Research Bureau, Health Canada, 420-757 West Hastings St. - Federal Tower, Vancouver, BC V6C 1A1, Canada. [2]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada. [3]Department of Public and Occupational Health, Amsterdam Public Health Research Institute, VU University Medical Center Amsterdam, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. [4]Barcelona Institute for Global Health (ISGlobal), 4c/ Rosselló, 132, 5th 2nd, 08036 Barcelona, Spain. [5]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [6]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [7]McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, ON, Canada.

Received: 19 April 2017 Accepted: 22 April 2017
Published online: 09 May 2017

References

1. European Commission, Directorate-General for Health and Consumers, Unit D3 eHealth and Health Technology Assessment. The use of big data in public health policy and research, background information document. European Commission: Brussels; 2014.
2. Khoury MJ, Ioannidis JPA. Big data meets public health: human well-being could benefit from large-scale data if large-scale noise is minimized. Science. 2014;346:1054–5. doi:10.1126/science.aaa2709.
3. Buhi ER. Digital health and AJPH: the time has come! Am J Public Health. 2015;105:420. doi:10.2105/AJPH.2015.302585.
4. Patrick K. Harnessing big data for health. CMAJ. 2016;188:555. doi:10.1503/cmaj.160410.
5. Malanga SE, Loe JD, Robertson CT, Ramos KS. Big data neglects populations most in need of medical and public Health Research and interventions. Arizona Legal Studies Discussion Paper. 2016:16–26.
6. Fawcett T. Mining the quantified self: personal knowledge discovery as a challenge for data science. Big Data. 2015;3:249–66. doi:10.1089/big.2015.0049.
7. Black ME. How data science will change public health. Thebmjopinion. 2015. http://blogs.bmj.com/bmj/2015/11/13/mary-e-black-how-data-science-will-change-public-health/. Accessed 12 Nov 2016.
8. Jumbe NL, Murray JC, Kern S. Data sharing and inductive learning—toward healthy birth, growth, and development. N Engl J Med. 2016;374:2415–7. doi:10.1056/NEJMp1605441.
9. Merson L, Gaye O, Guerin PJ. Avoiding data dumpsters–toward equitable and useful data sharing. N Engl J Med. 2016;374:2414–5. doi:10.1056/NEJMp1605148.
10. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, Heymann D, Burke DS. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014;14:1144. doi:10.1186/1471-2458-14-1144.
11. NIOSH. Exposome and Exposomics. 2016. http://www.cdc.gov/niosh/topics/exposome/. Accessed 20 Nov 2016.
12. Coughlin SS. Toward a road map for global -omics: a primer on –omic technologies. Am J Epidemiol. 2014;180:1188–95. doi:10.1093/aje/kwu262.
13. Manrai AK, Cui Y, Bushel PR, Hall M, Karakitsios S, Mattingly CJ, Ritchie M, Schmitt C, Sarigiannis DA, Thomas DC, Wishart D, Balshaw DM, Patel CJ. Informatics and data analytics to support exposome-based discovery for public health. Annu Rev Public Health. 2017;38:279–94. doi:10.1146/annurev-publhealth-082516-012737.
14. Stingone JA, Buck Louis GM, Nakayama SF, Vermeulen RCH, Kwok RK, Cui Y, Balshaw DM, Teitelbaum SL. Toward greater implementation of the exposome research paradigm within environmental epidemiology. Annu Rev Public Health. 2017;38:315–27. doi:10.1146/annurev-publhealth-082516-012750.
15. Turner MC, Nieuwenhuijsen M, Anderson K, Balshaw D, Cui Y, Dunton G, Hoppin JA, Koutrakis P, Jerrett M. Assessing the exposome with external measures: commentary on the state of the science and research recommendations. Annu Rev Public Health. 2017;38:215–39. doi:10.1146/annurev-publhealth-082516-012802.
16. Dennis KK, Marder ME, Balshaw DM, Cui Y, Lynes MA, Patti GJ, Rappaport SM, Shaughnessy DT, Vrijheid M, Barr DB. Biomonitoring in the era of the exposome. Environ Health Perspect. 2017;125:502–10. https://doi.org/10.1289/EHP474.
17. Starkey C, Garvin C. Knowledge from data in the built environment. Ann N Y Acad Sci. 2013;1295:1–9. doi:10.1111/nyas.12202.
18. Kitchin R. Big data, new epistemologies and paradigm shifts. Big Data Society. 2014:1–12. doi:10.1177/2053951714528481.
19. Institute of Medicine (US) Committee on Environmental Justice. Toward environmental justice: research, education, and health policy needs. Washington: National Academies Press (US); 1999.
20. NIH. Big data to knowledge (BD2K). 2017. https://datascience.nih.gov/bd2k. Accessed 23 Feb 2017.
21. European Commission. Big data supporting public health policies. 2016. https://ec.europa.eu/eip/ageing/funding/horizon-2020/big-data-supporting-public-health-policies-sc1-pm-18-2016_en. Accessed 20 Nov 2016.

22. European Commission. Environment and health. 2016. http://ec.europa.eu/research/health/index.cfm?pg=policy&policyname=environment. Accessed 20 Nov 2016.
23. HERCULES Exposome Research Center. 2016. http://emoryhercules.com/. Accessed 20 Nov 2016.
24. The Canadian Urban Environmental Health Research Consortium. 2016. http://canue.ca/. Accessed 20 Nov 2016.
25. Atkinson RW, Kang S, Anderson HR, Mills IC, Walton HA. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. Thorax. 2014;69:660–5. doi:10.1136/thoraxjnl-2013-204492.
26. Pope CA 3rd, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath CW Jr. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med. 1995;1513:669–74.
27. Dockery DW, Pope CA 3rd, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG Jr, Speizer FE. An association between air pollution and mortality in six U.S. cities. N Engl J Med. 1993;329:1753–9.
28. Pinault L, Tjepkema M, Crouse DL, Weichenthal S, van Donkelaar A, Martin RV, Brauer M, Chen H, Burnett RT. Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort. Environ Health. 2016;15:18. doi:10.1186/s12940-016-0111-6.
29. Crouse DL, Peters PA, Hystad P, Brook JR, van Donkelaar A, Martin RV, Villeneuve PJ, Jerrett M, Goldberg MS, Pope CA 3rd, Brauer M, Brook RD, Robichaud A, Menard R, Burnett RT. Ambient PM2.5, O_3, and NO_2 exposures and associations with mortality over 16 years of follow-up in the Canadian census health and environment cohort (CanCHEC). Environ Health Perspect. 2015;123:1180–6. doi:10.1289/ehp.1409276.
30. Künzli N, Jerrett M, Mack WJ, Beckerman B, LaBree L, Gilliland F, Thomas D, Peters J, Hodis HN. Ambient air pollution and atherosclerosis in Los Angeles. Environ Health Perspect. 2005;113:201–6.
31. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform. 2015;19:1216–23. doi:10.1109/JBHI.2015.2404829.
32. Licskai C, Sands TW, Ferrone M. Development and pilot testing of a mobile health solution for asthma self-management: asthma action plan smartphone application pilot study. Can Respir J. 2013;20:301–6.
33. OSHA. Improve Tracking of Workplace Injuries and Illnesses A Rule by the Occupational Safety and Health Administration on 05/12/2016 Federal Register. 2016;81 FR 29623.
34. Stocks SJ, McNamee R, van der Molen HF, Paris C, Urban P, Campo G, Sauni R, Martínez Jarreta B, Valenty M, Godderis L, Miedinger D, Jacquetin P, Gravseth HM, Bonneterre V, Telle-Lamberton M, Bensefa-Colas L, Faye S, Mylle G, Wannag A, Samant Y, Pal T, Scholz-Odermatt S, Papale A, Schouteden M, Colosio C, Mattioli S, Agius R, Working Group 2. Cost action IS1002—monitoring trends in occupational diseases and tracing new and emerging risks in a NETwork (MODERNET).. Trends in incidence of occupational asthma, contact dermatitis, noise-induced hearing loss, carpal tunnel syndrome and upper limb musculoskeletal disorders in European countries from 2000 to 2012. Occup Environ Med. 2015;72:294–303. doi:10.1136/oemed-2014-102534.
35. Wagner GR. Can predictive analytics help reduce workplace Risk? 2014. https://blogs.cdc.gov/niosh-science-blog/2014/10/02/pa/. Accessed 7 Dec 2016.
36. Rantonen O, Alexanderson K, Pentti J, Kjeldgard L, Hamalainen J, Mittendorf-Rutz E, Kivimäki M, Vahtera J, Salo P. Trends in work disability with mental diagnoses among social workers in Finland and Sweden in 2005-2012. Epidemiol Psychatri Sci (in press).
37. Schuring M, Robroek SJ, Otten FW, Arts CH, Burdorf A. The effect of ill health and socio economic status on labor force exit and re-employment: a prospective study with ten years follow-up in the Netherlands. Scand J Work Environ Health. 2013;39:134–43.