View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Explorer

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Exploiting big data for critical care research

Citation for published version:
Docherty, A & Lone, N 2015, 'Exploiting big data for critical care research' Current Opinion in Critical Care, vol. 21, no. 5. DOI: 10.1097/MCC.0000000000000228

Digital Object Identifier (DOI):
10.1097/MCC.0000000000000228

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Current Opinion in Critical Care

Publisher Rights Statement:
This is the author's final peer-reviewed manuscript as accepted for publication.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 05. Apr. 2019

Exploiting big data for critical care research

Introduction

Over recent years the digitalisation of data collection and storage, in combination with advances in data science, has enabled the linkage and analysis of large datasets, or ‘big data’. The proposed benefits of these large linked databases to health are significant, with faster progress in improving health, better value for money and higher quality science. This is of particular benefit for critical care research where conditions are rare, heterogeneous, with high mortality rates and high loss to follow-up. However, there are limitations in using data to answer questions different from its original purpose. In addition to this, there are ethical concerns regarding the confidentiality and autonomy of the patients who contribute to these datasets. In this review we will firstly define big data, and then explore the main sources of data collection, consider the uses for big data, discuss the limitations and ethical concerns of using these large datasets, and finally consider scope for future projects.

Definitions and sources of big data

There is no specific definition for big data, but it is generally understood to refer to datasets whose size, complexity and dynamic nature are beyond the scope of traditional data collection and analysis. The size of collected data may be the size of a petabyte (10^{15} bytes), but more usefully can be classified into the five ‘V’s: volume, velocity, variety, veracity, and value[1]. These ‘V’s represent the vast volumes, the high-speed real-time data, and the variety of data types from unstructured text to physiological data, to imaging and genomic sequencing. The analysis of such complex data requires methods more familiar to the field of informatics than clinical research, such as machine learning and computational linguistics[2].

Healthcare registries and databases

National critical care audit registries capture population-level critical care activity in participating hospitals with the primary aim of producing comparative risk adjusted outcomes in order to benchmark performance. The Intensive Care National Audit and Research Centre (ICNARC) has developed the Case Mix Programme, an audit of patient outcomes in England, Wales and Northern Ireland which contains 1.5 million patients and has formed the basis for clinical audit and research[3]. Other example registries include the Scottish Intensive Care Society Audit Group (SICSAG - Scotland)[4] and the Australian and New Zealand Intensive Care Society (ANZICS)[5] databases.

Health care databases maintained for administrative purposes, such as monitoring hospital activity or charging for health care provision, provide a rich source of data to identify patients requiring critical care or to link to critical care registries. In Scotland, all hospital discharges since 1981 are recorded in the Scottish Morbidity Record 01 (SMR01) database with complete national coverage[6]. Through use of standardised national patient identifiers, this dataset is routinely linked to the SICSAG registry and death records.

Electronic medical records

Primary and secondary care records have become increasingly computerised in order to facilitate the coordination of healthcare professionals working across different sites. As such they are a detailed account of each individual patient's interaction with the healthcare system. Data entry is increasingly standardised, facilitating linkage and research across different areas of healthcare. However unstructured 'free text' data can also be used for research purposes, and analysis techniques such as natural language processing (NLP) make this much more readily analysable[7].

There are several systems now in routine use in critical care that collate physiological data from bedside monitors, laboratory results, medical and nursing interventions and drug prescriptions onto a single electronic patient record[8]. These realtime data can be linked to other datasets to identify individual responses to interventions or physiological insults, increasing the accuracy of predictive models.

Laboratory and genetic data

There is a wealth of biomarker and imaging data that is collected as part of routine clinical care. Imaging analysis techniques have been developed that can automate data extraction, for example the detection of pulmonary emboli[9], which can then be linked to other patient datasets. Whole genome sequencing, epigenetics and genome wide association studies have also produced vast quantities of data to add to individual patient analysis.

Clinical trial data

The number of randomised controlled trials conducted in critical care has grown substantially over the past decade[10]. Information is meticulously collected during these trials, often to answer a single research question, and at substantial expense. The concept of 'open data' has been promoted to encourage sharing of these datasets to enable secondary analysis. This could yield considerable benefits, particularly when datasets from multiple studies are pooled. Individual level meta-analysis of trials is a self-evident application of this approach[11]. Furthermore, exploiting pooled existing trial data to answer novel, epidemiological research questions can result in the generation of new knowledge. However, additional energy and expense may be required to harmonise datasets by mapping variables and outcomes across datasets to allow analyses to be undertaken.

The TBI-IMPACT (International Mission for Prognosis and Analysis of Clinical Trials in Traumatic Brain Injury) project is an example of the step change in knowledge made possible through pooling clinical trial data[12]. This international collaborative project pooled trials relating to traumatic brain injury in order to develop innovative methods to improve trial design and analysis, refine outcomes, and improve risk prediction. To date, the collaboration has published over 60 peer-reviewed articles.

Examples of big data in critical care research

The potential benefits of these large and near-complete datasets for research are significant. These data are collected routinely, and research results generated from these sources would be available more quickly and at considerably less expense than prospective studies. Big data can refine questions, generate hypotheses, and identify potential recruits for experimental studies. It can also estimate the sample size for studies more accurately by modelling estimated event rates and treatment effects based on large retrospective data. PCORnet (Patient Centred Outcomes Research net) is an initiative to combine electronic health records and other electronic sources of very large populations, with patient-researcher partnerships[13]. Patients will have an active role by generating questions, sharing data, volunteering for interventional trials, and interpreting and disseminating results.

Genomics

Genome wide association studies compare the genetic variation of patients with the disease or trait (cases) and patients without (controls). Research in critical care genomics has demonstrated that human susceptibility to infection is strongly heritable[14]. Rautenan et al recently identified common variants in the FER (Fps/Fes related tyrosine kinase) gene that associate with a reduced risk of death from sepsis due to pneumonia[15]. This could enable us to tailor interventions based on the likelihood of individual patient response.

Health services research and audit

The increasing complexities of therapy available to us in critical care have been accompanied by increasing expenses, and it is important to justify these costs in terms of outcomes and efficiencies. Analysis of national and international critical care registries have enabled validation of advanced ICU scoring systems, prediction of diagnostic outcomes, decision support, and comparative analyses for the contributing hospital[3].

The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database consists of 25,328 intensive care unit stays[16]. It has collected clinical data, pooled from intensive care information systems and hospital archives, and high resolution physiological data (waveforms and time series of derived physiological measurements) obtained from bedside monitors. This has been de-identified for use as a freely accessible database for healthcare researchers.

Quality improvement

The Mayo Clinic is conducting a trial in Critical Care using ProCCESs AWARE - Patient Centered Cloud-based Electronic System: Ambient Warning and Response Evaluation[17]. This quality improvement tool combines electronic medical records, medical informatics and warning systems derived from data on previous patients. It will monitor adherence to best practice (eg appropriate shock resuscitation, appropriate sepsis treatment), and collect patient outcomes such as length of intensive care unit stay, and healthcare utilisation.

It is also possible to use these data streams in real time to detect adverse events: Thommandram et al describe the use of real time data to detect, analyse and classify neonatal cardiorespiratory spells, thus shortening the time to diagnosis[18].

Public health

Large datasets can identify disease patterns with the hope of targeting interventions at an earlier stage. Social media has been used with varying success in defining the geographical and temporal course of disease outbreaks, real-time tracking of harmful and infectious diseases, and increasing the knowledge of global distribution for various diseases[19,20].

Limitations

Having more data to analyse does not obviate the need to apply epidemiological conceptual reasoning coupled with a careful approach to study design. Chance, error, bias and confounding-fundamental concepts in epidemiology - must be considered in the design and analysis of all observational studies. In fact, analyses using big data may amplify these limitations through the false security of producing highly precise results with narrow confidence limits.

Chance

Big data should reduce the risk false negative results arising from analyses - a typical limitation of small, prospective studies. However, false positive results may be more likely to occur due to testing of multiple hypotheses and the ability to analyse numerous subgroups (‘data dredging’). Study registration and protocol publication *a priori* may reduce the risk of publication of spurious study results[21].

Error

Big data are usually collected for a purpose other than research. This means data can be compromised with respect to quality, accuracy, completeness, and accessibility (measurement error). Data science methods and machine learning algorithms are often employed to help make detect data signals amidst the background noise of messy data, but they cannot reliably remedy other data quality issues, particularly if researchers have a poor grasp of the nature of these limitations. The National Hip Fracture Database (NHFD) in England was mistakenly thought to contain 65,535 patients – this was in fact the maximum number of rows that could be processed by older versions of Microsoft Excel[22].

Bias and confounding

Of the numerous biases that can operate in observational studies, selection bias is of particular importance. The assumption that big data are representative of the target population to which results are generalised needs to be carefully assessed when relying on data generated through routine internet usage or social media. Studies using this population may exclude older individuals and those from poorer backgrounds – the very population groups usually over-represented in critical care.

A common pitfall of epidemiological analyses is the assumption that a correlation identified between two factors can be attributed to a causal relationship. The recent publicity surrounding 'Google Flu Trends' (GFT) highlighted this and other issues with big data analyses[23]. GFT used queries posed to Google search engine related to influenza, correlated with official influenza incidence estimates, to produce an algorithm which appeared to accurately predict influenza activity over a number of seasons[20]. However, the algorithm substantially overestimated influenza activity in 2012/2013 and subsequent seasons. This was in part attributable to the confounding effect of media attention on that season's influenza outbreak, which influenced the public to search for influenza-related items for reasons other than being ill with influenza symptoms. This episode emphasises the need for an understanding of fundamental epidemiological concepts when analysing big data[23].

Ethical and legal issues

When compared with the potential for harm in interventional trials in critical care, harm arising from research using big data is less obvious. However, there are two main ethical areas of concern: the potential for identifying individual patients and the invasion of privacy without consent.

Research using fully anonymised big data is consistent with Article 8 of the Human Rights Act 1998: 'everyone has the right to respect for his private and family life'. However linked databases are pseudonymised – they contain personal confidential data that has been anonymised but has a residual risk of re-identification. Linkage across multiple datasets increases the risk of individual re-identification, particularly for rare conditions. This may be of particular relevance in genome sequencing where patients who have traits associated with susceptibility to disease may have care rationed or withdrawn[24].

Insistence on formal consent for big data research could cause wider societal harm, as the participation bias which might arise could skew the data to such an extent as to make results inaccurate or meaningless[25]. In the specific context of prospective studies, concerns have been highlighted regarding sharing datasets to enable secondary analysis. Usually, study participants consent for their data to be used for a single, specific purpose. Consent taken from individuals for future secondary uses of the data is unlikely to be truly informed – the original researcher taking consent is unlikely to know the nature of research to be carried out in the future. In order to promote responsible data sharing, two guidance documents have recently been published both of which provide a framework to meet these ethical concerns[26,27].

The ethical basis for accessing and using big data for research without consent depends on societal expectations relating to how data will be used. Public concerns often centre around the disclosure and subsequent misuse of information to employers or insurance companies[28]. Workshops aimed at public engagement for population databases have shown a broad level of support for secondary use of data for research[28]. The National Health Service (NHS) in England argues for assumed individual patient consent for research using pseudonymised data combined with public awareness campaigns, but with an 'opt-out' option available. This allows patients a choice, thereby respecting autonomy and privacy and countering accusations of paternalism.

Safeguards are an essential component of maintaining data security and public confidence, and every organisation should put policies, procedures and systems in place to ensure confidentiality rules are followed[29]. In the UK, legislation allows for the imposition of heavy fines and ultimately prosecution for breaches of the law. In addition to legal scrutiny, it is essential that there is overview of research using big data by Institutional Review Boards or Research Ethics Committees. Use of accredited 'safe-havens' (restricted environments for the secure analysis of data), supported by robust protection and governance, is one method of maintaining public confidence in big data research[30]. Data are only made available to approved researchers once pseudonymisation has taken place and research proposals have been scrutinised for risks of disclosure.

Future

The use of big data in critical care research and clinical practice will continue to grow, driven by increasing availability of large clinical datasets, genetic databases, linkage to non-health care data and the flourishing collaborations with informatics experts and data scientists. The increasing use of social media as a healthcare tool means that interactions between patients, patient groups and clinicians can be integrated into research. Development of open source software, such as the statistical package 'R' (www.r-project.org), along with web-based interfaces for data entry which increase ease of adapting databases, will lead to data analysis becoming more accessible.

The potential to combine big data with 'small data' such as those from the numerous randomised controlled trials in critical care raises the tantalising possibility of individualised medicine. At the very least, these novel methods may help to identify more homogeneous subgroups for which treatments might not be effective or even harmful[31].

In order to move beyond the hyperbole surrounding big data, clinicians need to gain familiarity with the analytical techniques used by data scientists and, similarly, data scientists need to understand epidemiological concepts. Until then, clinicians and researchers should retain a healthy scepticism in the face of claims that big data circumvents the problems inherent in observational studies.

Conclusion

The critical care community are well placed to capitalise on the potential benefits of big data through the significant volumes of data collected from many sources. Big data will enable us to refine our descriptive and predictive analytics based on real data. Whilst not obviating the need for clinical trials, by defining questions more precisely and estimating more accurate sample sizes and event rates, it will enable us to improve their design. As big data becomes increasingly mainstream, it will be important to safeguard issues such as data security, governance and confidentiality. Big data has the potential to improve medical care and reduce costs significantly, both by individualising medicine, and by bringing together multiple sources of data about each individual patient.

References

1. Demchenko Y, Zhao Z, Grosso P, Wibisono A, de Laat C. Addressing Big Data Challenges for Scientific Data Infrastructure. *IEEE Computing Society*. 2012:614–617.

2. Gandomi, A. and M. Haider, Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 2015. 35(2): p. 137-144.

3. Harrison D a, Brady AR, Rowan K. Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit & Research Centre Case Mix Programme Database. *Critical care (London, England)*. 2004;8(2):R99–111.

4. SICSAG. Scottish Intensive Care Society Audit Group Annual Report: Audit of Intensive Care Units in Scotland 2014 Reporting on 2013. 2014 [Accessed on 08/04/2015]; Available from: http://www.sicsag.scot.nhs.uk/docs/SICSAG-report-2014-web.pdf?1.

5. ANZICS. Australia and New Zealand Intensive Care Society Annual Report: 2014. [Accessed 29/04/15]; Available from: http://www.anzics.com.au/Downloads/ANZICS%20Annual%20Report%202014.pdf

6. (http://www.adls.ac.uk/nhs-scotland/general-acute-inpatient-day-case-smr01/?detail).

7. FitzHenry F, Murff HJ, Matheny ME, et al. Exploring the Frontier of Electronic Health Record Sruveillance: The Case of Post-Operative Complications. *Med Care*. 2013;51(6):509–516.

8. CareVue: available at: http://www.healthcare.philips.com/pwc_hc/main/shared/ assets/documents/patient_monitoring/icip/icip_concept_wp_10.pdf

9. Masutani Y, Macmahon H, Doi K. Computerized Detection of Pulmonary Embolism in Spiral CT Angiography Based on Volumetric Image Analysis. *IEEE Trans Med Imaging*. 2002;21(12):1517–1523.

10. Michael O. Harhay, Jason Wagner, Sarah J. Ratcliffe, Rachel S. Bronheim, Anand Gopal, Sydney Green, Elizabeth Cooney, Mark E. Mikkelsen, Meeta Prasad Kerlin, Dylan S. Small, and Scott D. Halpern "Outcomes and Statistical Power in Adult Critical Care Randomized Trials", American Journal of Respiratory and Critical Care Medicine, Vol. 189, No. 12 (2014), pp. 1469-1478

11. Intensive Care Med (2010) 36:11–21

12. Maas AIR, Murray GD, Roozenbeek B, Lingsma HF, Butcher I, McHugh GS, Weir J, Lu J, Steyerberg EW, for the International Mission on Prognosis Analysis of Clinical Trials in Traumatic Brain Injury

(IMPACT) Study Group (2013) Advancing care for traumatic brain injury: findings from the IMPACT studies and perspectives on future research. Lancet Neurology, 12 (12): 1200-1210

13. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby J V, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association : JAMIA*. 2014;21(4):578–82.

14. Baillie JK. Targeting the host immune response to fight infection. *Science*. 2014;344:807–808.

15. Rautanen A, Mills TC, Gordon AC, et al. Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study. *The Lancet. Respiratory medicine*. 2015;3(1):53–60.

16. Saeed M, Villarroeol M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Crit Care Med*. 2011;39(5):952–960.

17. ProCCESs AWARE - Patient Centered Cloud-based Electronic System: Ambient Warning and Response Evaluation. Available at: https://clinicaltrials.gov/ct2/show/NCT02039297). [Accessed 29/04/2015]

18. Thommandram A, Pugh JE, Smieee JME, Smieee CM, James AG. Classifying Neonatal Spells Using Real - Time Temporal Analysis of Physiological Data Streams : Algorithm Development. *IEEE PHT*. 2013:240–243.

19. Broniatowski DA, Paul MJ, Dredze M (2013) National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. PLoS ONE 8(12): e83672. doi:10.1371/journal.pone.0083672

20. Ginsberg, J., et al., Detecting influenza epidemics using search engine query data. Nature, 2009. 457(7232): p. 1012-4.

21. Loder, E., T. Groves, and D. MacAuley, Registration of observational studies. 2010. 340: c950.

22. White S, Moppett I, Griffiths R. Big Data and Big Numbers. *Anaesthesia*. 2014;69(4):389–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24641648. Accessed February 24, 2015.

23. Lazer, D., et al., The Parable of Google Flu: Traps in Big Data Analysis. Science, 2014. 343(6176): p. 1203-1205.

24. Char DS, Cho M, Magnus D. Whole genome sequencing in critically ill children. *The Lancet. Respiratory medicine*. 2015;2600(15):6–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25704991. [Accessed 29/04/2015].

25. The Academy of Medical Sciences. *Personal data for public good : using health information in medical research*. London; 2006. Available at: www.acmedsci.ac.uk. [Accessed 29/04/2015]

26. Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials. Tudur Smith C, Hopkins C, Sydes M, Woolfall K, Clarke M, Murray G, Williamson P. April 2015

27. Institute of Medicine. Discussion Framework for Clinical Trial Data Sharing: Guiding Principles, Elements, and Activities. Washington, DC: The National Academies Press, 2014.

28. The Scottish Government. *Public acceptability of Data Sharing Between the Public, Private and Third Sectors for Research Purposes*.; 2013. Available at: http://www.scotland.gov.uk/Publications/2013/10/1304/8.

29. Caldicott F, Manning K. *A guide to confidentiality in health and social care.* London; 2013. Available at: http://www.hscic.gov.uk/media/12822/Guide-to-confidentiality-in-health-and-socialcare/pdf/HSCIC-guide-to-confidentiality.pdf.

30. Department of Health. *Information: To share or not to share? The Information Governance Review.*; 2013. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf .

31. Issa J.Dahabreh David M.Kent 2014. Can the Learning Health Care System Be Educated With Observational Data? JAMA Vol 312(2) 129,