Invited Commentary on Stewart and Davis " 'Big data' in mental health research – current status and emerging possibilities"

Jonathan D. Hafferty(1), Daniel J. Smith(2), Andrew M. McIntosh(1)

Affiliations

(1)Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
(2)Institute of Health and Wellbeing, University of Glasgow, Gartnavel Royal Hospital, Glasgow, UK

*Corresponding author : Jonathan Hafferty (email: jonathan.hafferty@ed.ac.uk)

Recent years have witnessed a revolution in data science and 'big data' into which psychiatric research has also being drawn. 'Big data' has been defined as data sets which are so large in size, so fast to change, and so complex in structure, that traditional data processing techniques are overwhelmed. [1] The mining and exploitation of such big data resources as Electronic Healthcare Records (EHRs) presents an exciting challenge to the field of psychiatric epidemiology. The number of big data projects within psychiatric research are growing and Stewart and Davis's literature review is therefore timely. [2]

Technological advances in data processing and storage, computer networking, mobile technology, and data manipulation, have rendered huge quantities of healthcare related data potentially amenable to analysis. Such studies potentially offer much larger patient numbers, wider parameters of study, and longer timescales of follow-up, than is typical of randomised controlled trials (RCTs) or cohort studies. A further advantage of this data is it often arises from naturalistic clinical settings, in terms of both clinical practice and patient health and comorbidity. Indeed, while often considered the 'gold standard' in medical research, RCTs do have important limitations, such as overly-strict exclusion criteria. Routine clinical datasets can therefore be complementary to RCT data, while also making research findings more relevant to everyday clinical practice. In addition, the quantity of clinical 'big data' potentially allows analysis of rarer clinical conditions, or subject areas that would be unlikely to meet ethical approval for more conventional studies (for example medication usage in pregnancy). Big data studies provide the scale and breadth of patient numbers required for stratified, predictive and personalised medicine research. Additionally, notwithstanding the many challenges of working with big data, large scale analysis of routinely collected healthcare datasets has already demonstrated effectiveness in fields such as pharmacovigilance [3] and post-marketing clinical trials.

As described in detail by Stewart and Davis, the challenging aspects of working with 'big data' are captured by the taxonomy of 'Vs' – volume, velocity and variety – first described by Laney [4] and extended since. [1,2] There are additional issues to working with healthcare big data specifically, including that clinical administrative data is generally not collected, curated or formatted in a manner optimised for research, and the inherent sensitivity of the data in terms of personal privacy. As the possibilities for healthcare big data expand into such areas as text mining and natural language processing of clinical records; near 'real time' updating of repositories; and incorporation of new data streams from mobile and wearable technology, robust systems must be in place to channel the potential data deluge. How great a role 'cloud' computing and storage will have in this is an intriguing question. In healthcare big data, the benefits offered by the cloud in increased power and accessibility, and reduced cost, must be weighed against concerns over data ownership, encryption and unauthorized access. An additional question is whether unique patient identifiers (UPIs) utilised by some national healthcare systems can be extended more widely, thereby facilitating securer forms of record linkage and de-identification as data is combined.

As with any relatively young scientific field, it is important that health informatics does not oversell its early potential. Google Flu Trends is a recent example of a machine-learning 'big data' approach which has fallen short of initial promise. [1, 5] Even within more traditional analysis of administrative clinical data, there are concerns about data quality and validity: examples including handling of missing data; the accuracy of DSM/ICD coding in clinical records [6]; and the use of psychiatric medication (which has multiple indications) as a surrogate for psychiatric caseness. Further research in these areas is required to guide good practice.

An important, and arguably under-researched, aspect of healthcare 'big data' is ethics and governance. Privacy, informed consent, data stewardship and the long-term ownership of data by academic and commercial entities, are becoming ever more pertinent issues as the pace of data accumulation increases. Data collected from individuals with psychiatric illness may come with additional consent and privacy concerns. It is important, however, that legitimate concerns about privacy do not inadvertently create an excessively restrictive regulatory environment. There is a critical role for policymakers in striking the right balance between privacy and realising the research potential of big data, as was recently illustrated in the debate regarding the European Data Protection Regulation. [7]

Furthermore, the expertise required for high quality big data analysis, ranging from data security to statistical modelling and machine learning, increases the scope for partnerships between commercial organisations specializing in these skills and clinical and academic healthcare entities. A recent example is the collaboration of Google Deepmind with Moorfields Eye Hospital, where anonymised access to

patient records was granted. [8] While such commercial-clinical-academic partnerships are a welcome development for the field, it is vital that they take place under a strong ethical and research governance framework to ensure clinician and patient confidence. [7]

As Stewart and Davis state, the future of big data research in psychiatry requires greater involvement of the research community in shaping the structure and content of clinical data platforms, rather than just being their passive recipients. This will require greater commonality of purpose and priorities between researchers, clinicians and patients. In order for EHRs and other administrative data to be better optimised for the requirements of large scale psychiatric research, GPs, clinical psychiatrists, hospital managers and patients will need to feel sufficiently motivated that it is in their interest to make the change. One way to achieve this is to ensure that research priorities align more closely with clinical priorities, such as outcomes-focused research.

As health informatics and data science develop as disciplines, they should seek to replicate advances that have proved advantageous in other fields. These include international collaboration, which has proven transformative in psychiatric genetics. The adoption of shared standards in data management, data governance and data security would facilitate joint, comparative and replicative big data studies between centres (for a summary of current initiatives see [9]). Similarly, the successful collaborations with patients and patient advocacy groups that have been a feature of clinical trials research provides a template for 'big data' as it seeks to persuade (sometimes sceptical) policymakers and the wider public of its potential.

There is also an urgent need to improve the awareness and training of academic psychiatrists regarding data science methods and this should be further developed. Raising the profile of 'big data' in healthcare research among primary and secondary care clinicians, and among other decision-makers within healthcare should be an essential aspect of the communication efforts of those engaged in the big data field in psychiatry.

As we look to the future, it is gratifying to note a number of important healthcare 'big data' initiatives taking shape. UK Biobank is an early exemplar of combining phenotypic, genetic, imaging and wearable technology data, with linkage to diverse primary and secondary care records, within an integrated platform that is available to researchers worldwide at a relatively low cost. The US Precision Medicine Initiative (PMI), announced last year by President Obama, is an attempt to do this on an even larger scale. Within psychiatry, the move towards near real-time access to clinical data incorporated in the SLaM BRC Clinical Record Interactive Search (CRIS) [10] will hopefully be among the first of many such resources. In addition, we are witnessing a broadening of the horizons of clinical record linkage, to encompass linkage to birth cohorts, genetic and imaging

repositories, but also to other administrative data in social, economic and educational fields. [2] This will allow future studies to better model the 'bio-psycho-social' outcomes of psychiatric illness for the benefit of patients, researchers and policymakers. We anticipate further, even more imaginative, uses of 'big data' resources, such as analysis of social media activity in psychiatric research, while recognising the potential biases as well as opportunities therein.

The volume, velocity and variety of big data is already far beyond that which can be comprehended and analysed solely by the human mind. Future developments as outlined above will render this even more so. In response to this, further innovation of data science techniques - machine learning, parallel computing, distributed data analysis, cloud based analysis and storage - is fundamental to the future of psychiatric health informatics. Familiarity with this toolbox of data science techniques will become a critical attribute for the psychiatric researcher of the 21st century. It should also be central to the strategies of psychiatric research funders.

While outlining the opportunities and challenges of 'big data' research in psychiatry, Stewart and Davis' review is also a call to the research community to play a greater role in shaping the development of these resources. We welcome and endorse this perspective.

Word Count : 1442 words

REFERENCES

[1] Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ (2015) Big data for health. IEEE J Biomed Health Inform. 19(4):1193-208.

[2] Stewart R, Davis K. (2016) 'Big data' in mental health research : current status and emerging possibilities. Soc Psychiatry Psychiatr Epidemiol. (8):1055-72.

[3] Harpe SE. (2009) Using secondary data sources for pharmacoepidemiology and outcomes research. Pharmacotherapy.(2):138-53

[4] D. Laney (2011) 3D data management: Controlling data volume, velocity and variety. Meta Group Inc., Stamford, CT, USA, Tech. Rep. 949

[5] Lazer D, Kennedy R, King G, Vespignani A. (2014) Big data. The parable of Google Flu: traps in big data analysis. Science. ;343(6176):1203-5.

[6] Monteith S, Glenn T, Geddes J, Bauer M. (2015) Big data are coming to psychiatry : a general introduction. Int J Bipolar Disord. 3(1):21

[7] Mittelstadt BD, Floridi L (2016) The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Sci Eng Ethics.;22(2):303-41.

[8] https://deepmind.com/health

[9] Jensen PB, Jensen LJ, Brunak S. (2012) Mining electronic health records : towards better research applications and clinical care. Nat Rev Genet. 13(6):395-405.

[10] Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, Fernandes A, Hayes RD, Henderson M, Jackson R, Jewell A, Kadra G, Little R, Pritchard M, Shetty H, Tulloch A, Stewart R. (2016) Cohort profile of the South London and Mausley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 1;6(3)