Data science for mental health: a UK perspective on a global challenge

Article *in* The Lancet Psychiatry · October 2016

DOI: 10.1016/S2215-0366(16)30089-X

Data science for mental health: a UK perspective on a global challenge

Andrew M McIntosh, Robert Stewart, Ann John, Daniel J Smith, Katrina Davis, Cathie Sudlow, Aiden Corvin, Kristin K Nicodemus, David Kingdon, Lamiece Hassan, Matthew Hotopf, Stephen M Lawrie, Tom C Russ, John R Geddes, Miranda Wolpert, Eva Wölbert, David J Porteous, for the MQ Data Science Group

Data science uses computer science and statistics to extract new knowledge from high-dimensional datasets (ie, those with many different variables and data types). Mental health research, diagnosis, and treatment could benefit from data science that uses cohort studies, genomics, and routine health-care and administrative data. The UK is well placed to trial these approaches through robust NHS-linked data science projects, such as the UK Biobank, Generation Scotland, and the Clinical Record Interactive Search (CRIS) programme. Data science has great potential as a low-cost, high-return catalyst for improved mental health recognition, understanding, support, and outcomes. Lessons learnt from such studies could have global implications.

Lancet Psychiatry 2016; 3: 993–98

Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK (Prof A M McIntosh MD, Prof C Sudlow PhD, Prof S M Lawrie MD, T C Russ PhD); **Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK** (Prof R Stewart PhD, K Davis MRCPsych, Prof M Hotopf PhD); **Swansea University Medical School, Swansea University, Swansea, UK** (A John MD); **Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK** (Prof D J Smith MD); **Department of Psychiatry & Psychosis Research Group, Trinity College Dublin, Dublin, Ireland** (A Corvin PhD); **Faculty of Medicine, University of Southampton, Southampton, UK** (Prof D Kingdon MD); **Health eResearch Centre, University of Manchester, Manchester, UK** (L Hassan PhD); **Department of Psychiatry, University of Oxford, Oxford UK** (Prof J R Geddes MD); **Child Outcomes Research Consortium (CORC) and Evidence Based Practice Unit, University College London, and Anna Freud Centre, London, UK** (M Wolpert DClinPsych); **Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK** (K K Nicodemus PhD, D J Porteous PhD); **and MQ Transforming Mental Health, London, UK** (E Wölbert PhD)

What is data science?

Data science is the extraction of knowledge from high-volume datasets by use of computing science and statistics (figure 1).[1] Data science is ubiquitous in global business and modern living and can partner technical revolutions such as those for medical genomics and imaging, to revolutionise the monitoring, diagnosis, treatment, and prevention of disease. These ideals are implicit in the fields of stratified medicine and precision medicine. The case for data science is often made for oncology, cardiology, and infectious diseases. Here we argue for the enormous potential of data science to transform mental health research and clinical practice worldwide. International collaboration will be necessary for maximum reach and impact. We review the available resources, barriers, and opportunities from a UK perspective before setting out how the full potential of data science could be realised on a global scale.

Why mental health and why now?

Mental disorders are arguably the greatest hidden burden of ill health, with substantial long-term impacts on individuals, carers, and society.[2] People with mental illness are often socially excluded[3] and are less likely to participate in research studies or remain in follow up.[4–6] Complexities of defining diagnoses present particular challenges for mental health research. Richly annotated, longitudinal datasets matched to data science analytics offer an unprecedented opportunity for more robust diagnostics, and also the prediction of outcome, treatment response, and patient preferences to inform interventions.[7] This might also provide more effective targeting of recruitment to observational and interventional studies. Such data are large in size and dimensions and require the application of advanced analytics, such as machine learning, whereas more conventional techniques are less computationally tractable.

A key issue in data science is the description of data types that are the most informative, readily available, and efficiently captured. Generic data types include electronic health and prescribing records, education, welfare, sociodemographic, laboratory, and real-world monitoring through wearable devices and environmental sensors. More specific data might include genomic data, in-vivo brain imaging, and cognitive traits. Important challenges include shortcomings in dataset completeness and linkage potential, as well as acceptability to patients and the wider public given the perceived sensitivity of mental health data. It is also important to consider the types of information that can create new ways of classifying mental health and illness and be universally applied beyond the perfect-world discovery setting.

UK resources to pioneer this approach

Population cohorts

Several UK population cohorts have enhanced clinical, biological, and social datasets linked to routinely collected electronic data. The UK Biobank is a cohort study of half a million individuals aged 37–73 years recruited between 2006 and 2010. Participants completed a touch-screen questionnaire, underwent an interview, and participated in several assessments including measures of depressive symptoms, distress, cognition, and alcohol and cigarette use. Additionally, linkages have been made to National Health Service (NHS) health-care episode data, and a number of biological measures taken, including DNA for whole-genome genotyping. An initial pilot medical imaging study included collection of unprocessed brain structure, function, and connectivity data from more than 5000 participants; this is now being extended to 100 000 individuals. Further longitudinal and outcome assessments include repeat cognitive testing and

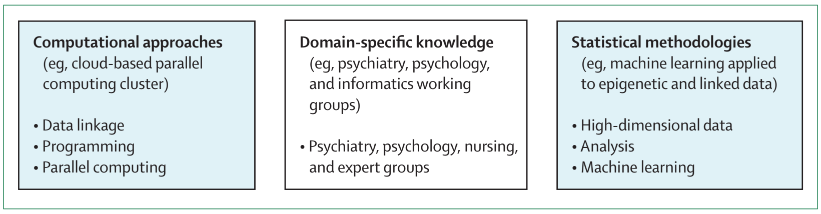
Computational approaches
(eg, cloud-based parallel computing cluster)
• Data linkage
• Programming
• Parallel computing
Domain-specific knowledge
(eg, psychiatry, psychology, and informatics working groups)
• Psychiatry, psychology, nursing, and expert groups
Statistical methodologies
(eg, machine learning applied to epigenetic and linked data)
• High-dimensional data
• Analysis
• Machine learning

Figure 1: Components of data science

Correspondence to:
Prof Andrew M McIntosh,
Division of Psychiatry,
University of Edinburgh,
Royal Edinburgh Hospital,
Edinburgh EH10 5HF, UK
andrew.mcintosh@ed.ac.uk

actigraphy. Lifetime history of mental illness will be assessed in greater depth with a web-based questionnaire. UK Biobank thus brings unprecedented deep and broad phenotyping to mental health research.[8]

For more on the **UK Biobank** see http://www.ukbiobank.ac.uk

Several other UK population cohorts have done detailed phenotyping of participants and have the potential for record linkage to routine health-care and administrative data. Notable examples include the Generation Scotland: Scottish Family Health Study (GS:SFHS),[9,10] a family and population based study located in Scotland with near complete record linkage; and the Avon Longitudinal Study of Parents and Children,[11] a UK-based cohort study with data available from before birth to more than 20 years follow-up. Further information on these and other studies is provided in the appendix.

See Online for appendix

Domain-specific cohorts linked to routinely collected data

By contrast with population-based research cohorts, several UK resources are focused on mental health and routinely collect clinical data from the NHS, the UK's comprehensive health-care provider. These data might be more representative of the general population and provide a framework for implementation. The National Centre for Mental Health (NCMH) was established in Wales in 2011, and partnered to the Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics. The NCMH recruits participants with mental disorders to the NCMH cohort, currently more than 6000 individuals, who are willing to participate in research and be recontacted. Clinical data (eg, demographic characteristics, routine secondary care, expanded clinical, neuropsychological, and imaging information) and biological samples are collected to create a platform and infrastructure for mental health research into the causes and treatment of mental illness and learning disability. In 2015, the Farr Institute (founded in 2013 with the aim of harnessing health data for patient and public benefit by facilitating the safe and secure use of electronic patient record and other population-based datasets) partnered with the NCMH allowing for linkage of the cohort to routine data nested within prevalent diagnostic electronic cohorts within the Secure Anonymised Information Linkage (SAIL) databank.[12,13] The SAIL databank is a whole-population- anonymised health record from roughly 3·5 million patients describing primary care, hospitals, child health, education, cause-specific mortality, deprivation, and urbanicity. Participants can be tracked across health and social care settings, while also protecting privacy in accordance with relevant legislation using a split-file approach.[12,13] This is the first time that genomic data have been linked to the SAIL databank,[14] allowing researchers to address questions on the impact of genetic, environmental, and health factors including modifiable lifestyle factors on clinically important outcomes.

For more on the **National Centre for Mental Health** see http://www.ncmh.info

For more on the **Secure Anonymised Information Linkage (SAIL)** see http://www.saildatabank.com

Electronic health record-derived cohorts

The increasing use of electronic health records is creating databases unparalleled both in sample size and in the depth of information contained. The use of these data for research is encouraged by policy[15,16] and subject to necessary technical and ethical considerations.[17–19] An important distinction is made between structured information and unstructured text—the former being simpler to analyse, although clinical uncertainties are often poorly coded.[20–23] Here, text mining can be used alongside structured information to better define groups.[24,25] Structured information about patients needing specialist care has been collected systematically by the NHS since 1981, through Hospital Episode Statistics in England, the Scottish Morbidity Record, and Patient Episode Data for Wales. These structured records are available to researchers as linked data and are published in open-access aggregated form,[26,27] along with primary care data.[1,28] Despite concerns about the speed and accuracy of these electronic data,[29,30] they might prove valuable to measure real-world outcomes and assess their mediators and predictors.

The Clinical Record Interactive Search (CRIS) application was developed at the South London and Maudsley NHS Foundation Trust in 2007, as a means of rendering the large volumes of electronic mental health record data available for research.[31,32] CRIS accesses mental health case records from around 260 000 patients within a south London geographic catchment area containing approximately 1·2 million residents; replications of CRIS have recently become operational elsewhere in London, Oxford, and Cambridge. Key to the development are data structuring and deidentification pipelines and a wider data security and governance model which has been patient led from the outset.[33] Research applications have included searches to help identify and characterise rare scenarios for further investigation,[34,35] and data linkage projects to characterise physical health outcomes.[36,37] Recent enhancements include the development of natural language processing applications to derive structured information from the text fields present in electronic mental health records. These include recorded diagnoses, cognitive test scores, pharmacotherapy, and symptom profiles.[38–42] The Child Outcomes Research Consortium approach is a flagship UK electronic record project and is based around the outcomes of children and adolescents seen in specialist mental health services (appendix).[43]

Linkage to real-time health data and wearable devices

Companies such as Apple (HealthKit and ResearchKit) and Google (Alphabet) are developing health-based applications and wearable devices, as part of a wider array of environmental sensors, the so-called internet of things, and health application developer toolkits. The potential to capture relevant real-time and longitudinal health data (eg, mood, diet, activity, and sleep patterns), matched to physiological measures (eg, heart rate, blood glucose, cortisol), is potentially transformative and pervasive at low cost and independent of conventional health-care provision. A good, early example of such an initiative in psychiatry is True Colours), a platform developed to capture continuous patient-generated data with the required usability and acceptability to permit reliable longitudinal follow-up. Notably, this technology is being piloted as a supplement to routine health care.

Public trust and clinical governance

Although UK data science resources provide major opportunities to improve research and health services, they demand public support, public trust, and transparent governance arrangements. The MRC Farr Institute, the European Data in Health Research Alliance, and Patients4Data have all promoted the importance of data sharing for research and health-care impact while acknowledging the potential risks of inaccurately recorded information and data breaches.

Attitudes research suggests that mental health data are among the most personal and sensitive.[44,45] There are diverse reasons why people might be reluctant or unwilling to consent to the use of their data for mental health research.[46,47] Encouragingly, studies indicate that most mental health service users agree to the use of their health records for research, particularly when efforts to engage in ongoing communication about their use and potential benefits are made.[32,48] It is important to reflect how cancer research has largely dispelled the past stigma of a cancer diagnosis—can modern day research, driven by data science, do the same for mental health? We think so, by reframing and redefining the causes and by reshaping and revitalising effective interventions.

Safe and transparent models of governance for reuse of mental health data are essential to maintain public trust. Systems have been developed that protect privacy and, in the future, innovations such as dynamic models of consent[49] might also allow the public further control over their data. The Farr Institute includes a programme of public engagement with a focus on the safe and transparent use of patient and research data.

The Scottish model is a useful example of how data science and record linkage can be done at scale and in a trusted environment. Scotland has excellent administrative and health-care data resources. The NHS Community Health Index (CHI), a unique personal identifier for 99% of the population, has greatly enabled pseudonymised linkage between health and administrative data (figure 2). Arguably, the other key to unlocking the benefits of routinely collected data in Scotland has been the presence of good research governance procedures and proactive engagement with the public to drive forward health informatics research. Public input in review of grant applications is standard practice, and includes providing lay research summaries and wider dissemination in addition to public consultation and outreach. Consultation work suggests that the public supports the use of administrative and health data in research, provided there is adequate data security and access is limited to personnel conducting research for public benefit. The public appears more supportive of academic and clinical research than it does of work conducted by commercial organisations.[44,51] All data outputs are scrutinised to ensure they do not identify individuals or breach privacy before being released. Open access summaries are published online as a condition of all research. Support to researchers throughout this process is provided by an eData Research and Innovation Service.[50]

For more on **True Colours** see https://oxfordhealth.truecolours.nhs.uk/www/en

Training, resource, and capacity implications

Challenges to be overcome

The availability and development of excellent resources for data science alongside robust governance procedures are necessary prerequisites for good data science. We argue that there are also specific technological and skills challenges to be overcome and that fulfilling the promise of data science will involve international collaboration spanning high-income and low-income countries.

For more on the **European Data in Health Research Alliance** see http://datasaveslives.eu

For more on **Patients4Data** see https://twitter.com/patients4data

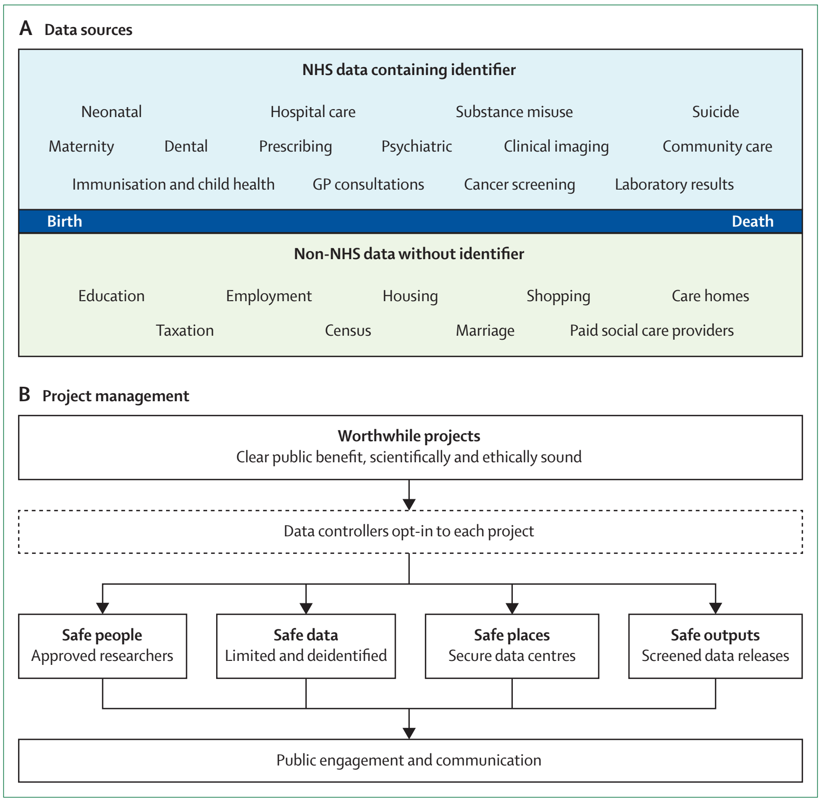

Figure 2: The Scottish model

Shows the linkable data sources available in Scotland (A), whose linkage is facilitated by the unique Community Health Index number and project management of data science projects (B). NHS=National Health Service. GP=general practitioner. Reproduced with permission of Pavis and Morris.[50]

Technological resource

The capacity of data storage and access, and the personnel to collect and analyse data, are rate limiting steps in the ongoing development of data science. Routinely collected administrative and health data tend to be financed centrally by government, but have limited phenotypic coverage and have, until recently, been used mainly for planning. More detailed phenotyping is possible with use of routine clinical data, such as CRIS in London and Psychosis Clinical Information System (PsyCIS) in Glasgow,[52] and large-scale genetic, -omics, and neuroimaging studies generate huge volumes of data that pose tractable data storage issues. The combination of these datasets is very challenging and requires data harmonisation and compatibility issues to be addressed.

Databases need to gather and hold data, and enable users to search for and access data of interest to them. Data sharing agreements and how to facilitate collaboration and innovation are key issues for data scientists. In practice, data-generation projects decide on a case-by-case basis what they will offer to centralised depositories without proposing a coordinated solution for how the data will be linked to other sources. Centralised databases can be more attractive to data depositors by offering managed data access and trusted analysis environments. The Global Alliance for Genomics and Health brings together different health sectors and regions worldwide, to catalyse the sharing of methods to harmonise data approaches across diverse datasets.

For more on the **Global Alliance for Genomics and Health** see https://genomicsandhealth.org

For more on **MQ** see https://www.mqmentalhealth.org

Skills resource

Identifying, training, and fostering a generation of clinically informed data scientists from a wide range of backgrounds should be a top priority. For this, multidisciplinary training programmes are needed that will expose scientists, informaticians, and statisticians to commonly used clinical data, diagnoses, and treatments, as well as a range of relevant methodological approaches. Data scientists will usually need further postgraduate training in statistics and computational methods. Trainees will need to be familiar with ethical and regulatory requirements, and be prepared to become familiar with the diverse ways in which health data are recorded and stored. Given the diversity of resources and methodologies, a variety of approaches seems inevitable. Particular care and attention to the career structure of data scientists will be needed to nurture early career researchers and ensure that expensively acquired expertise is not lost after training. A spectrum of skills and disciplines needs to be present in a data science team and its leadership as well as a common understanding of the need for complementary expertise. As data science evolves in fields such as engineering and finance, there will be opportunities to learn from their experience.

Search strategy and selection criteria

Individual studies, projects and databases were identified by each of the authors. Exemplars providing good examples of data, methods or projects in a specific area of Data Science were agreed by consensus. Information presented from each study and topic area was agreed with at least one additional author.

National and international collaboration

To achieve maximum reach and impact, development and maintenance is needed of international and interdisciplinary databases and the networks to support their efficient use. There are many examples of this process working well in areas such as genomics[53] and brain imaging,[54] where international consortia have brought together databases of unparalleled size and scope. There are particular challenges in expanding these initiatives to low-income and middle-income countries where the infrastructure might be poor and low-cost methods of data collection and storage will be needed. Clinical information from paid and public health providers might also come with differing governance frameworks and commercial interests, but overcoming these barriers will prove beneficial for all parties.

There is much work to be done in standardising assessments, outcome measures, and terminology within, let alone between, countries. UK and international research charities such as MQ, the Wellcome Trust, and publicly funded research councils have an important role to play in matching researchers and their research questions to datasets spanning multiple subject domains and countries. Routine health record data with detailed mental health coverage are stored in parts of the UK, Australia, and the exemplary Scandinavian systems. Some projects, such as the UK Biobank, encourage external data analysis even as data are being collected, whereas others will not be openly shared until the original funder-approved aims have been met. Subject to regulatory approvals, systems should be put in place to facilitate the incorporation of data from time-limited projects as soon as is practicable. Intellectual property and resource considerations could make this incorporation challenging. Fostering collaborations, developing safe havens to facilitate joint working, and convening advisory groups with wide representation will help to enhance complementarity across projects and data collections.

Our vision of the future

Against a backdrop of no fundamentally new pharmacological treatment in the past 60 years and a progressive pharmaceutical industry withdrawal from mental health research and development, an alternative course is essential. Mental health remains the leading area of unmet medical need in the developed world and it is rapidly acquiring the same status in the developing world.

Combining large health-care and administrative datasets with real-time monitoring, laboratory, genomic, and imaging data could achieve a step-change in the way health care is provided and research is organised. In our opinion, data science will greatly enhance our ability to conduct discovery science, epidemiological studies, personalised medicine, and plan services. Without the better understanding of mental health problems that will come with use of big data, longer-term visions for self-management, better treatments, and learning health systems will not be possible. It is thus vital that current initiatives in data science recognise and support this need.

Contributors

AMM proposed the manuscript and its contents. All authors participated in the MQ Data Science meeting contributing to its planning and intellectual content. AMM and DJP cowrote the first draft of the manuscript. LH wrote the public trust and clinical governance section. AC, DS, and EW contributed to the literature search. All authors contributed to editing and revising the manuscript, AMM incorporated the edits of the other authors. All authors approved the final version.

Declaration of interests

AMM has received research funding from Pfizer, Janssen, and Eli Lilly. RS has received research funding from Pfizer, Lundbeck, Roche, Janssen, and GlaxoSmithKline. SML has received grants and personal fees from Roche, grants from Pfizer and Abbvie, and personal fees from Janssen and Sunovion. All other authors declare no competing interests.

Acknowledgments

AMM, SML, and DP are supported by the Wellcome Trust through a strategic award (reference 104036/Z/14/Z). The manuscript was critically revised by Gerard Leavey (University of Ulster, UK); Graham Moon (University of Southampton, UK); Rosie Cornish (University of Bristol, UK); Tamsin Ford (University of Exeter Medical School, UK); Gary Donohoe (NUI Galway, UK); Rudolf Cardinal (University of Cambridge, UK); Zina Ibrahim (King's College London, UK); Margaret Maxwell and Nadine Dougall (University of Stirling, UK); Felicity Callard (University of Durham, UK); and David McDaid (London School of Economics and Political Science, UK).

References

1 Dhar V. Data Science and Prediction. *Commun Acm* 2013; **56:** 64–73.

2 Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* 2013; **382:** 1575–86.

3 Barr B, Kinderman P, Whitehead M. Trends in mental health inequalities in England during a period of recession, austerity and welfare reform 2004 to 2013. *Soc Sci Med* 2015; **147:** 324–31.

4 van Heuvelen MJG, Hochstenbach JBM, Brouwer WH, et al. Differences between participants and non-participants in an RCT on physical activity and psychological interventions for older persons. *Aging Clin Exp Res* 2005; **17:** 236–45.

5 Rogers A, Harris T, Victor C, et al. Which older people decline participation in a primary care trial of physical activity and why: insights from a mixed methods approach. *BMC Geriatr* 2014; **14:** 46.

6 Goldberg M, Chastang JF, Zins M, Niedhammer I, Leclerc A. Health problems were the strongest predictors of attrition during follow-up of the GAZEL cohort. *J Clin Epidemiol* 2006; **59:** 1213–21.

7 Torous J, Baker JT. Why psychiatry needs data science and data science needs psychiatry: connecting with technology. *JAMA Psychiatry* 2016; **73:** 3–4.

8 Smith DJ, Nicholl BI, Cullen B, et al. Prevalence and characteristics of probable major depression and bipolar disorder within UK biobank: cross-sectional study of 172,751 participants. *PLoS One* 2013; **8:** e75362.

9 Smith BH, Campbell A, Linksted P, et al. Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* 2012; **42:** 689–700.

10 Smith BH, Campbell H, Blackwood D, et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* 2006; **7:** 74.

11 Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013; **42:** 97–110.

12 Ford DV, Jones KH, Verplancke JP, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009; **9:** 157.

13 Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009; **9:** 3.

14 Lloyd K, McGregor J, John A, et al. A national population-based e-cohort of people with psychosis (PsyCymru) linking prospectively ascertained phenotypically rich and genetic data to routinely collected records: overview, recruitment and linkage. *Schizophr Res* 2015; **166:** 131–6.

15 Department of Health. Personalised health and care 2020: using data and technology to transform outcomes for patients and citizens. London: HM Government, 2014.

16 Clarke A, Adamson J, Sheard L, Cairns P, Watt I, Wright J. Implementing electronic patient record systems (EPRs) into England's acute, mental health and community care trusts: a mixed methods study. *BMC Med Inform Decis Mak* 2015; **15**: 85.

17 Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *J Intern Med* 2013; **274:** 547–60.

18 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Rev Genet* 2012; **13:** 395–405.

19 Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. London: Nuffield Council on Bioethics, 2015.

20 Morrison Z, Fernando B, Kalra D, Cresswell K, Sheikh A. National evaluation of the benefits and risks of greater structuring and coding of the electronic health record: exploratory qualitative investigation. *J Am Med Inform Assoc* 2014; **21:** 492–500.

21 Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FD. Envisioning a learning health care system: the electronic primary care research network, a case study. *Ann Fam Med* 2012; **10:** 54–59.

22 Bernat JL. Ethical and quality pitfalls in electronic health records. *Neurology* 2013; **81:** 1558.

23 Whooley O. Diagnostic ambivalence: psychiatric workarounds and the Diagnostic and Statistical Manual of Mental Disorders. *Sociol Health Illn* 2010; **32:** 452–69.

24 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013; **20:** 144–51.

25 Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol* 2012; **8:** e1002823.

26 Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes-systematic review and recommendations for reporting. *Eur J Public Health* 2013; **23:** 86–92.

27 Health & Social Care Information Centre. Users and Uses of Hospital Episode Statistics. Leeds: Health & Social Care Information Centre, 2012.

28 Heath & Social Care Information Centre. Supporting open data and transparency. Leeds: Health & Social Care Information Centre, 2015.

29 RSA Open Public Services Network. Exploring how available NHS data can be used to show the inequality gap in mental healthcare. London: RSA Open Public Services Network, 2015.

30 CAPITA. The quality of clinical coding in the NHS. London: CAPITA, 2014.

31 Perera G, Broadbent M, Callard F, et al. Cohort profile of the South London and Maudsley 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* 2016; **6:** e008721.

32 Stewart R, Soremekun M, Perera G, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009; **9:** 51.

33 Fernandes AC, Cloete D, Broadbent MT, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak* 2013; **13:** 71.

34 Su YP, Chang CK, Hayes RD, et al. Retrospective chart review on exposure to psychotropic medications associated with neuroleptic malignant syndrome. *Acta Psychiatr Scand* 2014; **130:** 52–60.

35 Oram S, Khondoker M, Abas M, Broadbent M, Howard LM. Characteristics of trafficked adults and children with severe mental illness: a historical cohort study. *Lancet Psychiatry* 2015; **2:** 1084–91.

36 Chang CK, Hayes RD, Perera G, et al. Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London. *PLoS One* 2011; **6:** e19590.

37 Chang CK, Hayes RD, Broadbent MT, et al. A cohort study on mental disorders, stage of cancer at diagnosis and subsequent survival. *BMJ Open* 2014; **4:** e004295.

38 Patel R, Jayatilleke N, Broadbent M, et al. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method. *BMJ Open* 2015; **5:** e007619.

39 Patel R, Lloyd T, Jackson R, et al. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ Open* 2015; **5:** e007504.

40 Perera G, Khondoker M, Broadbent M, Breen G, Stewart R. Factors associated with response to acetylcholinesterase inhibition in dementia: a cohort study from a secondary mental health care case register in london. *PLoS One* 2014; **9:** e109484.

41 Kadra G, Stewart R, Shetty H, et al. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. *BMC Psychiatry* 2015; **15:** 166.

42 Hayes RD, Downs J, Chang CK, et al. The effect of clozapine on premature mortality: an assessment of clinical monitoring and other potential confounders. *Schizophr Bull* 2015; **41:** 644–55.

43 Fleming I, Jones M, Bradley J, Wolpert M. Learning from a Learning Collaboration: The CORC Approach to Combining Research, Evaluation and Practice in Child Mental Health. *Adm Policy Ment Health* 2014; **43:** 297–301.

44 Wellcome Trust. Qualitative Research into Public Attitudes to Personal Data and Linking Personal Data. London: Wellcome Trust, 2013.

45 Taylor MJ, Taylor N. Health research access to personal confidential data in England and Wales: assessing any gap in public attitude between preferable and acceptable models of consent. *Life Sci Soc Policy* 2014; **10:** 15.

46 Ridgeway JL, Han LC, Olson JE, et al. Potential bias in the bank: what distinguishes refusers, nonresponders and participants in a clinic-based biobank? *Public Health Genomics* 2013; **16:** 118–26.

47 Papoulias C, Robotham D, Drake G, Rose D, Wykes T. Staff and service users' views on a 'Consent for Contact' research register within psychosis services: a qualitative study. *BMC Psychiatry* 2014; **14:** 377.

48 Callard F, Broadbent M, Denis M, et al. Developing a new model for patient recruitment in mental health services: a cohort study using Electronic Health Records. *BMJ Open* 2014; **4:** e005654.

49 Williams H, Spencer K, Sanders C, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. *Jmir Med Inf* 2015; **3:** e3.

50 Pavis S, Morris AD. Unleashing the power of administrative health data: the Scottish model. *Public Health Res Pr* 2015; **25:** e2541541.

51 Willison DJ, Steeves V, Charles C, et al. Consent for use of personal information for health research: Do people with potentially stigmatizing health conditions and the general public differ in their opinions? *BMC Med Ethics* 2009; **10:** 10.

52 Martin DJ, Park J, Langan J, Connolly M, Smith DJ, Taylor M. Socioeconomic status and prescribing for schizophrenia: analysis of 3200 cases from the Glasgow Psychosis Clinical Information System (PsyCIS). *Psychiatr Bull* 2014; **38:** 54–57.

53 O'Donovan MC. What have we learned from the Psychiatric Genomics Consortium. *World Psychiatry* 2015; **14:** 291–93.

54 Thompson PM, Stein JL, Medland SE, et al. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav* 2014; **8:** 153–82.