

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

Andrew M McIntosh^{*,1}, Robert Stewart², Ann John³, Daniel J Smith⁴, Katrina Davis², Cathie Sudlow¹, Aiden Corvin⁵, Kristin K Nicodemus¹⁰, David Kingdon⁶, Lamice Hassan⁷, Matthew Hotopf², Stephen M Lawrie¹, Tom C Russ¹, John R Geddes⁸, Miranda Wolpert⁹, Eva Wölbert¹¹, David J Porteous¹⁰, and the MQ Data Science Group¹¹

Affiliations

- 1 Division of Psychiatry, University of Edinburgh, Edinburgh, UK
- 2 King's College London (Institute of Psychiatry, Psychology and Neuroscience), London, UK
- 3 Swansea University Medical School, Swansea University, Swansea, UK
- 4 Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK
- 5 Department of Psychiatry & Psychosis Research Group, Trinity College Dublin, Dublin, Ireland
- 6 Faculty of Medicine, University of Southampton, Southampton, UK
- 7 Health eResearch Centre, University of Manchester, Manchester, UK
- 8 Department of Psychiatry, University of Oxford, Oxford UK
- 9 Evidence Based practice Unit, University College London, and Anna Freud Centre, London, UK
- 10 Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
- 11 MQ Transforming mental health, London, UK

Corresponding author*

Abstract	148 words
Main text	3435 words
Collaborating authors	155 words
Tables	0
Figures	2

Abstract

Data science extracts new knowledge from high dimensional datasets through computer science and statistics. Mental health research, diagnosis and treatment can benefit from data science data integration from consented cohort studies, genomics, routine healthcare and administrative data. The UK is well placed to do so through exemplary data science projects, such as UK Biobank, Generation Scotland, ALSPAC and CRIS, set within the UK National Health Service. The NHS ensures the highest standards of data governance, public trust and value, establishing the ideal platform for the advancement of mental health through data science. Data science has great potential as a low cost, high return catalyst for how mental health problems may be better recognised, understood, supported and outcomes improved. The UK NHS and embedded cohort studies provide a unique opportunity to field trial this approach to mental health, with global reach in terms of both their output and impact.

WHAT IS DATA SCIENCE?

Data science is the extraction of knowledge from high-volume data, using skills in computing science, statistics and the specialist domain knowledge of experts.¹ Data science pervades global business and modern living and can partner technical revolutions, such as medical genomics and imaging, to revolutionise the monitoring, diagnosis, treatment and prevention of disease. This ideal of interdisciplinary, evidence-based, data-driven precision medicine is implicit in the research strategies of major funding organisations. The case for data science is often made for cancer, heart and infectious diseases and yet, mental disorders rarely feature as first line targets. Here we argue that it is time to recognise the enormous potential for data science to transform mental health research and clinical practice.

Figure 1: What is data science?

Insert Figure 1 here

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 these conditions are often socially excluded³ and less likely to participate in research studies or adhere to follow up.^{4–6} Complexities around defining diagnoses present challenges for research into mental health conditions and thus enhanced longitudinal datasets are needed to supplement observational research. Data science offers an unprecedented opportunity not only for more robust diagnostics, but also the prediction of outcome, treatment response, and patient preferences to inform interventions.⁷ It may 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 analytical techniques, such as machine learning, where more conventional techniques are less computationally tractable.

A key issue in data science is the description of data types that are the most informative, most readily available and most easily and efficiently captured. Generic data types include electronic health and prescribing records, education, welfare, socio-demographic, laboratory and real world monitoring through wearable devices and environmental sensors. More specific data might include genomic data, *in vivo* brain imaging and cognitive/behavioural/psychological traits. Important challenges should be recognised, which 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, in the round, create new ways of classifying mental health and illness.

WHAT RESOURCES DO WE HAVE IN THE UK?

Within the UK we have rich resources for data science in mental health. In this rapidly moving field, we have selected a few exemplars for data science in the UK. Whilst these examples give an indication of the range of current resources, they also provide a template from which to shape the profile of data science for mental health in future years.

1. Population cohorts

There are several UK population cohorts with enhanced clinical, biological and social datasets linked to routinely collected electronic data. UK Biobank (www.ukbiobank.ac.uk) and Generation Scotland (www.gen.scot) are two examples which illustrate the range of possibilities available.

UK Biobank

UK Biobank is a cohort study of 502,000 individuals aged between 37 and 73 years who were recruited from 2006 to 2010. Participants attended an assessment centre, completed a touch-screen questionnaire, underwent a nurse-led interview, and took part in several assessments. These included measures of depressive symptoms, psychological distress, cognitive function and alcohol and cigarette use. In addition, linkages have been made to National Health Service (NHS) healthcare episode data, and a number of biological measures have been taken, including DNA for whole-genome genotyping. An initial pilot medical imaging study includes unprocessed brain structure, function and connectivity

data in over 5,000 participants, which is in the process of being extended to 100,000 individuals. Further longitudinal and outcome assessments include repeat cognitive testing and 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

Generation Scotland

The 'Generation Scotland: Scottish Family Health Study' (GS:SFHS) is a community and family based study of ~24,000 participants aged between 18 and 98. Participants were recruited from ~7,000 family groups^{9,10} with approval for medically relevant research, including genetic studies, and for re-contact. Questionnaire-based family history and demographics; detailed clinical data; validated measures of pain, cognition and mental health; pedigree information; and biological samples are available for 21,516 participants. All GS participants have consented to allow linkage to their medical record including all routine Scottish Morbidity Records, NHS prescriptions, mortality records and the Scottish Birth Record. Genetic and phenotypic data are held separately and de-identified. Whole-genome data is also available, and serum and urinary proteomics studies are under development. All participants were screened for lifetime depression: 2,706 participants (13.5%) met the DSM IV criteria for major depressive disorder (MDD).¹¹ Participants are currently being re-contacted for questionnaire-based assessment and further clinical and brain imaging measures.

2. Domain specific cohorts linked to routinely collected data (NCMH and SAIL Databank)

In contrast to population based research cohorts, several UK resources are focussed specifically on Mental Health and routinely collected clinical data from the NHS, the UK's comprehensive healthcare provider. These data may be more representative of the general population and provide a framework for implementation.

NCMH and SAIL Databank

The National Centre for Mental Health (NCMH) was established in Wales in 2011

and supports and performs high quality research into the causes and treatment of mental illness and learning disability (www.ncmh.info). The NCMH continues to expand a lifespan-wide cohort of 6000 participants, crossing multiple diagnostic categories, willing to participate in research who have consented to further contact. Recruitment involves collection of participant information (e.g. demographic, clinical, neuropsychological and imaging), routine clinical NHS secondary care data and biological materials.

NCMH has subsequently developed a research platform and infrastructure for mental health research in Wales. The NCMH cohort can now be linked to routine data nested within prevalent diagnostic cohorts. These cohorts can be tracked across healthcare settings, whilst protecting privacy, through linkage to routine electronic health and social datasets from the Secure Anonymised Information Linkage (SAIL) databank (www.saildatabank.com).^{12,13}

The SAIL databank is a research data repository for Wales, holding over 2Bn anonymised health records from ~3.5M patients. Primary care data include demographics, diagnoses, symptoms, referrals, laboratory investigations and prescriptions. Linkage to the hospital inpatient database provides information on admissions, diagnoses, surgery, treatment and discharges. Coverage of outpatient appointments, emergency attendances, child health, educational attainment, cause-specific mortality, deprivation and urbanicity are also recorded. Use of data is conducted in accordance with data protection and information governance legislation using a split file approach.^{12,13}

3. Electronic health record derived cohorts and the Farr Institute

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^{14,15} subject to technical and ethical considerations.^{16–21}

An important distinction is made between structured information and unstructured text – the former being simpler to analyse, albeit that clinical uncertainties are often poorly coded.^{22–26} Here, text mining may need to be employed alongside structured information to better define groups.^{18,27}

Structured information on patients requiring 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 are available to researchers as linked-data and are published in open-access aggregated form,^{28,29} along with primary care data related to the Quality Outcomes Framework and Increasing Access to Psychological Therapies.³⁰ Despite concerns about the speed and accuracy of these data,^{31–33} these resources may prove valuable for measuring real-world outcomes and assessing their mediators and predictors.

In 2013 electronic medical record linkage was given further impetus by the founding of the UK Farr Institute for Health Informatics Research. It has the aim of harnessing health data for patient and public benefit by facilitating the safe and secure use of electronic patient records and other population based data sets.

CRIS

The Clinical Record Interactive Search (CRIS) application was developed at the South London and Maudsley NHS Foundation Trust (SLAM) in 2007 as a means of rendering the large volumes electronic mental health record data available for research.^{34,35} CRIS at SLAM accesses mental health case records from around 260,000 patients within a south London geographic catchment of approximately 1.2m residents; replications of CRIS have recently become operational elsewhere in London, Oxford and Cambridge. Key to the development are not only the data structuring and de-identification pipeline afforded by CRIS itself, but also the wider data security and governance model which has been patient-led from the outset.³⁶ Research applications have included searches to help identify and characterise rare scenarios for further investigation,^{37,38} and data linkage projects to characterise physical health outcomes.^{39,40} Recent enhancements include the development of natural language processing applications to derive structured information from the text fields present in the electronic mental health record. These include recorded diagnoses, cognitive test scores, pharmacotherapy and symptom profiles.^{41–46}

Child Mental Health: The Child Outcomes Research Consortium approach

The Child Outcomes Research Consortium (CORC) is a practice research network of >50% of all child mental health providers in the UK (~70 organisations) with collaborators in Scandinavia and Australia. Members of this not-for-profit collaboration share pseudonymised child-level data annually, which are held by CORC centrally and consist of ~250,000 care episodes over ten years. The collaboration includes health and education providers and the voluntary sector. There is an initiative supported by the Department of Health to support closer data linkage between these datasets in future.⁴⁷ The CORC approach is an example of both the opportunities and the challenges in collecting and using routinely collected data from mental health service providers, and the use of 'deep domain knowledge'. CORC is committed to use of the data to inform "Precision Mental Health"⁴⁸ whilst also being mindful of the complexities, limitations and flaws in the data.⁴⁹ CORC draws on this data to support clinical decision-making, performance management, quality improvement and specific research studies.

Linkage to 'real-time' health data and wearable devices

The increasing use of wearable devices, such as activity monitors, smartphones and watches has provided a vast new source of health data. One of the main advantages is the rapid availability of 'real-time' data (e.g. steps and sleep patterns), which can include contemporaneous measures of heart rate, mood, diet, sleep and biochemistry. Access to personally generated data can be provided through smartphone or web-based applications that collect data from individuals, analyse them and provide consent to external researcher's data requests. 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 Internet of Things', and health application developer toolkits. This field is at an early stage and there are good examples of such initiatives in psychiatry. For example, Truecolours (<https://oxfordhealth.truecolours.nhs.uk/www/en/>) is a platform that has been developed to capture continuous patient-generated data with the required usability and acceptability to permit reliable longitudinal follow-up. It is our opinion that wearables and real-time data have the potential to transform disease monitoring, the relationship between patients and their healthcare

providers and provide new insights into the phenotype and neurobiology of mental disorders.

PUBLIC TRUST AND CLINICAL GOVERNANCE

Government administrative and healthcare data represent major resources for research and health service improvement. However, public support, public trust and governance arrangements are fundamentally important if their full potential is to be realised. There is a need for researchers, clinicians and policy-makers to engage with patients and the public in discussions about the potential benefits of research and risks of identification or privacy breaches. Several organisations are leading projects in the UK, including the Farr Institute ([#datasaveslives](#)), the European Data in Health Research Alliance ([datasaveslives.eu](#)) and Patients4Data ([patients4data.co.uk](#)). These campaigns engage with the public and policymakers to promote the power of patient data and influence regulatory authorities to allow continuing use of data for patient and public benefit. These initiatives have the potential to shape research questions and ensure these are meaningful for communities and patient groups, and will help to ensure sustainability. Such activity also sets the work within a wider public context and will improve visibility, transparency, and acceptability.

Health data are personal and sensitive, and attitudes research suggests that mental health data are among the most sensitive.^{50,51} Existing initiatives demonstrate that substantive patient involvement from the start (in the design, implementation, use and development of data-sharing initiatives), can help to ensure that meaningful progress is made.³⁵ It is important to attend to the growing body of research that helps us to understand the diverse reasons why people might be reluctant or unwilling to consent to the use of their data for mental health research.^{52,53} Studies indicate, encouragingly, that a majority of mental health service users agree to the use of their health records for research – particularly when efforts to engage in on-going communication about their use and potential benefits are made.^{35,54} There may be lessons to be learned from cancer research, which has been transformed from being characterised by stigma and under-funding into a highly successful global research movement.

Safe and transparent models of governance for re-use of mental health data are essential for developing and maintaining public trust. Systems have successfully been developed that protect privacy whilst enabling research in the public interest. In the future, innovations that allow the public further control over their data may offer further opportunities, such as dynamic models of consent⁵⁵ and crowdsourcing (e.g. www.PatientsLikeMe.com). The recently established 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'

Scotland is known to have some of the best administrative and care data in the world. In recent years, Scotland has developed an approach which has successfully delivered a number of informatics projects involving academia, industry and health service providers. One major reason for this success is the Community Health Index (CHI) - a unique identifier for approximately 99% of the population, facilitating pseudonymised linkage between health and administrative data (Figure 1).

Figure 2. National level data resources in Scotland⁵⁶

Insert Figure 2 here

The 'Scottish model' is an exemplar of how to ensure trustworthy data governance and engage with the public to drive forward health informatics research. Since 2002, the Scottish Government have had an engagement group to ensure public input into activities such as reviewing grant applications, providing lay research summaries and wider dissemination activities. Consultation work since then suggests that the public is content to offer support for the use of administrative and health data in research, provided there are robust processes to ensure that data security and limited access to trusted personnel conducting research for public benefit. It appears that the public is more supportive of academic and clinical research than work conducted by commercial organisations.^{50,57}

All the outputs generated are scrutinised to ensure they do not identify individuals or breach privacy before being released. Plain English summaries of research are published online and open access publication is a condition of all research. Support to researchers throughout this process is provided by a 'research coordinator' within an eData Research and Innovation Service.⁵⁶ The key elements of the Scottish model are illustrated within Figure 2.

The role of medical research charities

The role of research charities in the evolution of data access and utilisation is still emerging. Recent activity, led by the Association of Medical Research Charities has been focused at the national and international level, advocating for clearer statutory guidelines on oversight and accountability for NHS England as well as ensuring access to data for research purposes in European Union Data Protection legislation. As facilitators of a UK-wide discussion of research opportunities and challenges in mental health data science, MQ is working with charities and government to ensure that mental health is represented in critical discussions.

TRAINING, RESOURCE AND CAPACITY IMPLICATIONS

1. Technological resource

The capacity of data storage and access, and the personnel to collect and analyse data (and financing to maintain these) are rate-limiting steps in the ongoing development of data science. Routinely-collected 'administrative' and health data tend to be centrally financed by government but have limited phenotypic coverage and have, until recently, been used mainly for planning. More detailed phenotyping is possible in routine clinical data, such as CRIS in London and PsyCIS in Glasgow,⁵⁸ 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 for 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. Agreements about data sharing and how to facilitate collaboration and innovation are key issues for data scientists. In practice, data generation projects are deciding on a case by case basis what they they will offer to centralised depositories without offering a coordinated solution for how that data will be linked to other sources. Centralised databases can make themselves more attractive to data depositors by offering managed data access and trusted analysis environments. Existing examples, focussed on genetics, include the Global Alliance for Genomics and Health (GA4GH <https://genomicsandhealth.org>) and RD-Connect (<http://rd-connect.eu>).

2. Skills resource

Identifying, training and fostering a generation of clinically-informed data scientists from a wide range of backgrounds must be a top priority. This requires multidisciplinary training programmes, which 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 modelling. All trainees will need to be familiar with ethical and regulatory requirements as well as 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 and desirable. 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.

3. National and international collaboration

There is a need to develop and maintain international and interdisciplinary databases and the networks to support their efficient use. There is much work to be done in standardising assessments, outcome measures and terminology within, let alone between, nations. MQ has recently established UK-based

research charity with international reach dedicated to mental health (www.joinmq.org). MQ and other research charities such as the Wellcome Trust and Medical Research Council 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 include those stored by the Information and Statistics Division of the Scottish Government, a similar resource in Australia and the exemplary Scandinavian systems. Some projects, like 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, it is desirable that systems should be put in place to facilitate the incorporation of data from time-limited projects as soon as practicable. Intellectual property and resource considerations may make this challenging. Fostering collaborations, developing safe havens to facilitate joint working and convening advisory groups with wide representation will help enhance complementarity across projects and data collections.

OUR VISION OF THE FUTURE

Against a backdrop of no fundamentally new pharmacologic 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 is rapidly acquiring the same status in the developing world.

Combining large healthcare and administrative datasets with real-time monitoring, laboratory, genomic and imaging data could achieve a step change in the way healthcare 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

The manuscript was critically revised by:

Gerard Leavey, The Bamford Centre for Mental Health and Wellbeing, University of Ulster, United Kingdom. Graham Moon, Geography and Environment, University of Southampton, United Kingdom. Rosie Cornish, School of Social and Community Medicine, University of Bristol, United Kingdom. Tamsin Ford, University of Exeter Medical School, Exeter , United Kingdom. Gary Donohoe, Center for Neuroimaging and Cognitive Genomics (NICOG), School of Psychology, NUI Galway, United Kingdom. Rudolf Cardinal, Department of Psychiatry, University of Cambridge, United Kingdom. Zina Ibrahim, Department of Social Genetic & Developmental Psychiatry Kings College London, United Kingdom. Margaret Maxwell, NMAHP Research Unit, School of Health Sciences, University of Stirling, Stirling, United Kingdom. Nadine Dougall, NMAHP Research Unit, School of Health Sciences, University of Stirling, Stirling, United Kingdom. Felicity Callard, PhD, Department of Geography, University of Durham, United Kingdom. David McDaid, Personal Social Services Research Unit, London School of Economics and Political Science, London WC2A 2AE, United Kingdom

Legend to Figure 1 What is data science:

Figure showing the component features of data sciences

Legend to Figure 2: The 'Scottish' Model

Figure shows the linkable data sources available in Scotland, whose linkage is facilitated by a unique identifier: the CHI number

Legend to Figure 3: How and where data science can improve psychiatric diagnosis and optimise care

Currently, diagnosis depends upon a self-reported check list following DSM and ICD criteria. The patient will normally present in primary care and be diagnosed by a GP. The care pathway recommended may vary from monitoring only, self-help, counselling or drug prescription. Only a minority of patients with a confirmed diagnosis will respond to the initially selected intervention. Severe cases will be referred to secondary care and the process refined, but a significant minority will remain treatment refractive after testing all available therapies. There are currently no robust methods to stratify patients within a given diagnostic category that supports treatment choice or predicts treatment outcome. Counselling is resource limited and costly compared with drug prescription and this influences health practitioner decision making. The figure highlights in bold how data science could provide robust methods to stratify patients within a given diagnostic category that can support treatment choice and predict treatment outcome (disease stratification and precision medicine). The approach builds upon and extends beyond the NIMH Research Domain Criteria (RDoC) principles (<http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml>). Indicative, non-exhaustive data sources are listed as inputs towards refined diagnostic classification, optimal choice of first line treatment, development of better treatments and monitoring of response. Implicit in this refined approach is that a) multiple relevant data types and sources are readily available at low cost, b) advance epidemiological, statistical, and machine learning data analytics can extract the maximally informative variables, c) the chronic and remitting course of mental illnesses highlights the added value of objective longitudinal and real-time data, for input and feedback to patients, practitioners and carers, d) individual personal predictions and care-pathway monitoring will benefit from similarity matching within patient cohorts and against population norms.

References

- 1 Dhar V. Data Science and Prediction. *Commun ACM* 2012; **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* 2015; **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* 2013; **42**: 689–700.
- 10 Smith BH, Campbell H. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability.

- BMC Med Genet* 2006; **7**.
- 11 Fernandez-Pujals AM, Adams MJ, Thomson P, *et al*. Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in Generation Scotland: Scottish Family Health Study (GS:SFHS). *PLoS One* 2015; **10**: e0142197.
 - 12 Ford D, Jones K. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009; **9**.
 - 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 National Information Board, Department of Health. Personalised health and care 2020: Using Data and Technology to Transform Outcomes for Patients and Citizens. HM Government, 2014 DOI:10.1177/0272989X06295361.
 - 15 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 Inf Decis Mak* 2015; **15**: 15–204.
 - 16 Häyrynen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int J Med Inform* 2008; **77**: 291–304.
 - 17 Spriggs M, Arnold M V, Pearce CM, Fry C. Ethical questions must be considered for electronic health records. *J Med Ethics* 2012; **38**: 535–9.
 - 18 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inf Assoc* 2013; **20**: 144–51.
 - 19 Coorevits P, Sundgren M, Klein GO, *et al*. Electronic health records: new opportunities for clinical research. *J Intern Med* 2013; **274**: 547–60.
 - 20 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012; **13**: 395–405.
 - 21 Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. 2015

- http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf.
- 22 Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FDR. Envisioning a learning health care system: the electronic primary care research network, a case study. *Ann Fam Med* 2012; **10**: 54–9.
 - 23 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 Informatics Assoc* 2014; **21**: 492–500.
 - 24 Bernat JLMD. Ethical and quality pitfalls in electronic health records. *Neurology* 2013; **80**: 1057–61.
 - 25 Eason K, Waterson P. Fitness for purpose when there are many different purposes: Who are electronic patient records for? *Health Informatics J* 2014; **20**: 189–98.
 - 26 Whooley O. Diagnostic ambivalence: psychiatric workarounds and the Diagnostic and Statistical Manual of Mental Disorders. *Sociol Health Illn* 2010; **32**: 452–69.
 - 27 Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol* 2012; **8**: 27.
 - 28 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 Heal* 2013; **23**: 86–92.
 - 29 Health & Social Care Information Centre. Users and Uses of Hospital Episode Statistics. 2012
http://www.hscic.gov.uk/media/10495/Users-and-uses-of-HES/pdf/HES_Users_and_Uses.pdf.
 - 30 Health & Social Care Information Centre. Supporting open data and transparency. 2015. <http://www.hscic.gov.uk/transparency>.
 - 31 RSA Open Public Services Network. Exploring how available NHS data can be used to show the inequality gap in mental healthcare. 2015.
 - 32 CAPITA. The quality of Mental Health care cluster costing and activity in the NHS. 2014

- http://www.chks.co.uk/userfiles/files/The_quality_of_Mental_Health_care_cluster_costing_and_activity_data_in_the_NHS.pdf.
- 33 CAPITA. The quality of clinical coding in the NHS. 2014
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/364476/The_quality_of_clinical_coding_in_the_NHS.pdf.
- 34 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*.
- 35 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.
- 36 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.
- 37 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.
- 38 Oram S, Khondoker M, Abas M, Broadbent M, Howard LM. Characteristics of trafficked adults and children with severe mental illness: a historical cohort study. *The lancet Psychiatry* 2015; **2**: 1084–91.
- 39 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.
- 40 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.
- 41 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.
- 42 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.
- 43 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.
- 44 Wu C-Y, Chang C-K, Hayes RD, Broadbent M, Hotopf M, Stewart R. Clinical risk assessment rating and all-cause mortality in secondary mental healthcare: the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) Case Register. *Psychol Med* 2012; **42**: 1581–90.
- 45 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.
- 46 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.
- 47 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 Heal Ment Heal Serv Res* 2014; : 1–5.
- 48 Bickman L, Lyon A, Wolpert M. Achieving precision mental health through effective assessment, monitoring, and feedback processes: introduction to the special issue. *Adm Policy Ment Heal Ment Heal Serv Res* DOI: 10.1007/s10488-016-0718-5.
- 49 Wolpert M, Deighton J, De Francesco D, Martin P, Fonagy P, Ford T. From 'reckless' to 'mindful' in the use of outcome data to inform service-level performance management: perspectives from child mental health. *BMJ Qual Saf* 2014; **0**: 1–5.
- 50 Wellcome Trust. Summary Report of Qualitative Research into Public Attitudes to Personal Data and Linking Personal Data. 2013 http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp053205.pdf.

- 51 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. 2014; : 1–24.
- 52 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.
- 53 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**: 1–8.
- 54 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–e005654.
- 55 Williams H, Spencer K, Sanders C. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. *JMIR Med Informatics* 2015; **3**: e3.
- 56 Pavis S, Morris AD. Unleashing the power of administrative health data: the Scottish model. *Public Heal Res Pract* 2015; **25**: e2541541.
- 57 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**. DOI: 10.1186/1472-6939-10-10.
- 58 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**: 54–7.