How data science can advance mental health research

Short title: Data science and mental health

Tom C. Russ,1-5* Eva Woelbert,6 Katrina A.S. Davis,7,8 Jonathan D. Hafferty,2
Zina Ibrahim,9,10 Becky Inkster,11 Ann John7, William Lee,12,13 Margaret Maxwell,14
Andrew M. McIntosh,1,2 Rob Stewart,7,8 and the MQ Data Science group6
1. Centre for Cognitive Ageing & Cognitive Epidemiology, University of Edinburgh
2. Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh
3. Centre for Dementia Prevention, University of Edinburgh
4. Alzheimer Scotland Dementia Research Centre, University of Edinburgh
5. Old Age Psychiatry, Royal Edinburgh Hospital, NHS Lothian
6. MQ: Transforming Mental Health, London
7. Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London
9. Department of Biostatistics and Health Informatics, King’s College London, London
10. The Farr Institute of Health Informatics Research, University College London, London
11. Department of Psychiatry, University of Cambridge
12. Community and Primary Care Research Group, Plymouth University Peninsula Schools of Medicine and Dentistry, University of Plymouth
13. Devon Partnership NHS Trust, Devon
14. University of Stirling

* Correspondence to: Dr Tom C. Russ, Division of Psychiatry, Kennedy Tower, Royal Edinburgh Hospital, Morningside Terrace, Edinburgh EH10 5HF UK.

Telephone: +44 (0)131 537 6672; Email: T.C.Russ@ed.ac.uk

Word count: 4281
ABSTRACT

Accessibility of powerful computers and availability of so-called “big data” from a variety of sources means that data science approaches are becoming pervasive. However, their application in mental health research is often considered to be at an earlier stage than in other areas despite the complexity of mental health and illness making such a sophisticated approach particularly suitable. In this article we discuss current and potential applications of data science in mental health research using the UK Clinical Research Collaboration classification: underpinning research; aetiology; detection and diagnosis; treatment development; treatment evaluation; disease management; and health services research. We demonstrate that data science is already being widely applied in mental health research, but there is much more to be done now and in the future. The possibilities for data science in mental health research are substantial.

[135 words]

Key words: Data science, mental health, data mining

Summary for journal homepage:

Russ et al. discuss the broad applications of data science to mental health research and consider future ways that big data could improve detection, diagnosis, treatment, health care provision, and disease management.
INTRODUCTION

Data-driven approaches have become key in answering important scientific questions. In the UK, major developments, including that of Health Data Research UK, have highlighted the potential of these approaches.¹ However, definitions of “data science” lack clarity. It has been variously defined as: the ‘fourth paradigm’ of science (with empirical, theoretical, and computational science); a concept to unify statistics, data analysis, and their related methods; a synonym of statistics; and as that activity containing multidisciplinary investigations, models, and methods for data. Unsurprisingly, there is similarly little consensus in the curricula of the various data science degrees.

However, the world is changing because of the deluge of data generated daily as well as the growing capability of computers. It is probably fair to describe data science as generating new knowledge from real world data. This is distinct from mere description but additionally means deriving meaningful inferences from which it is possible to make helpful decisions, for example about treatment for a particular individual.

Some areas of medicine are already benefitting from data science, such as prevention of heart disease,² and treatment of some cancers.³ For mental health and neuroscience, data science is at a relatively early stage. In this article we outline our view of what data science might offer mental health research (summarised in Figure 1). We use as a framework the Health Research Classification System of the UK Clinical Research Collaboration (http://www.hrcsonline.net/) which divides health research under the headings shown in Box 1.
UNDERPINNING RESEARCH

**Figure 2** summarises the interrelation between applications of data science in mental health and illness, outlined in the following sections. This research is underpinned by studies of normal biological, psychological and social processes. Mental disorders are often aetiologically complex, with multiple environmental, psychological and genetic contributors and putative individual risk factors. Generating study samples of sufficient size to investigate this complexity and capture normal trait variation, has until recently been a prohibitive challenge. In the years ahead, data science holds transformative potential for research, through exploitation of emerging “big data” resources. This should enable underpinning research to drive informative new models of pathophysiology and novel therapeutic strategies.

One crucial development towards fulfilling this potential has been the establishment of highly scaled, population-based, collaborative biobanks. These offer phenotyping of mental, cognitive, and socioeconomic attributes, alongside genetic, biochemical and imaging data. Research within UK Biobank (N=500,000), for example, has identified 148 independent genetic loci associated with variations in cognitive function, which has relevance to researchers trying to understand how cognitive processes can be impaired in mental disorders. Other examples include the US Million Veteran’s Program (current N=600,000) and the BioBank Japan project (N=200,000). The biobanking model itself is also driving innovation in mental health research through the founding of bespoke biobanks with an explicit focus on mental disorders, such as the University of Michigan Mental Health BioBank (https://medicine.umich.edu/dept/psychiatry/news/archive/201708/u-m-department-psychiatry-depression-center-launch-mental-health-biobank).
Translating candidate genetic loci from GWAS into credible models of function requires deeper understanding of underlying biological processes. Advances in sequencing technology, such as the long awaited $100 per individual next-generation techniques,\textsuperscript{10} will expand available genomic data by orders of magnitude. Mental health researchers will require skills in data science – parallel computing, highly scalable storage, complex multivariate analysis, and visualisation – to make these data tractable\textsuperscript{11}. Combined with the related discipline of bioinformatics, data science offers great potential for underpinning research through \textit{in silico} discovery science, such as machine-learning-driven modelling of cellular protein folding\textsuperscript{12} and synaptic transmission,\textsuperscript{13} or computational modelling of receptor pharmacology for drug development.\textsuperscript{14} In addition to further stimulating research on the molecular biology of psychiatric disorders, this work also supplements preclinical psychopharmacological research.\textsuperscript{15}

Similarly, progress in neuroimaging has been constrained by a lack of sufficiently powered and consistent datasets of healthy controls\textsuperscript{16} through which to understand healthy brain development and morphology. New networks of open-access, multicentre imaging consortia, containing very large numbers of participants drawn from the general population (e.g. ENIGMA [\(N=50,000\)]\textsuperscript{17}, UK Biobank [Target \(N=100,000\)]\textsuperscript{5}, and the Human Connectome Project\textsuperscript{18}) enable brain mapping at scale across the life course. Research using scans from childhood to adulthood have revealed normal brain functional networks (e.g. default-mode, salience, sensorimotor) with distinct developmental trajectories.\textsuperscript{19} Alteration in the default-mode network has been implicated in follow-up research in autism,\textsuperscript{20} schizophrenia\textsuperscript{21} and Alzheimer’s dementia\textsuperscript{22} among others.\textsuperscript{23} Looking to the future, combining multiband imaging with genomics to study the inter-relationships between genotype, brain structure and functional architecture\textsuperscript{24} is a petabyte-scale data science challenge which can now be undertaken.\textsuperscript{25}
One of data science’s most exciting opportunities is the adaption of emerging technologies to generate new phenotypic and biometric variables, which can be linked to existing datasets. The development and commercialisation of a number of sensors, wearables, and smartphone applications, will enable real-time, fine-grained, monitoring of a number of traits, including sleep and mood variation,\textsuperscript{26} vital signs variation,\textsuperscript{27} and alcohol use\textsuperscript{28}. As well as guiding research in areas such as circadian rhythms in mental illness,\textsuperscript{29} such tools can also be used to augment clinical care.\textsuperscript{30} Diverse phenotypic data can also be extracted through data mining techniques deployed to social media and health records databases.\textsuperscript{31,32} Finally, record linkage to repositories of social and educational data\textsuperscript{33} will enable construction of a truly integrative model of mental processes, in health as well as disease.\textsuperscript{34}

**AETIOLOGY, PREVENTION OF DISEASE AND PROMOTION OF WELL-BEING**

All mental disorders have a complex aetiology.\textsuperscript{35} As well as providing causal information and potentially helping to refine current phenotypic definitions, genetic data resources need to be combined with more detailed longitudinal data on the environment. A key challenge is therefore better cohort characterisation on a large scale using participant-active repeated sampling as well as longitudinal data linkage.

There is growing interest in the physical environment and its relevance to mental health. For example, there are benefits to living in proximity to, and spending time in, green space.\textsuperscript{36-38} Such environmental data are often collected at the regional or neighbourhood scale – for example weather or air pollution monitoring stations – but allocate a value for the exposure (e.g., air pollution) based on an individual’s residential address.\textsuperscript{39}
More ambitious studies have recently generated a more nuanced quantification of physical environments using GPS technology, allowing observational data to be used to model the extent to which the environment might facilitate health and maximise well-being. For example, by monitoring the routes people take through the environment, alterations could encourage time spent in green spaces, thus allowing people potentially to benefit. Whilst there are inaccuracies in the location derived from GPS, and in the availability of openly available data on the local area in which participants live, these data offer the exciting prospect of increased accuracy in relating the environment to mental health.

Other large scale and challenging environmental data such as latitude, sunlight exposure, and ambient temperature have also been studied in relation to mental health. A number of vitamins, including folic acid and vitamins B₁₂ and D have been linked with mental illnesses. Mendelian randomization has also been used to infer a causal relationship of lower vitamin D with Alzheimer’s disease. Without data science, none of these ambitious studies would be possible.

Socioeconomic position is an important confounder of most environmental risk factors. For example, living close to a major road is associated with lower socioeconomic position in most parts of the world which may explain observed associations with important mental health outcomes. Ascertaining socioeconomic position on a large scale often involves linking an individual’s residential address with an area-based level of deprivation (which may or may not match that individual’s socioeconomic status). Novel data linkages – linking health to survey data with better socioeconomic measures or even to robustly anonymised Census output – is one way forward. There might also be scope to identify educational and occupational status from electronic health and other records. However, this approach is currently accompanied by substantial practical and ethical challenges.
Linking large well-characterised datasets with other sources of outcome data, for example electronic medical records or mortality records, provides important opportunities for the study of mental health and disorder. Data linkages allow for passive follow-up of research participants and outcomes of interest to be collected for relatively low unit cost and low drop-out. This is the approach being taken in UK Biobank.\(^5\) Such linkages clearly offer more outcomes than could feasibly be assembled through original data collection. However, case finding in primary and secondary care is likely to identify different patient groups\(^5\) and many people with mental disorders are not in contact with clinical services – and may not have a diagnosis and would not be identified by this approach.\(^5\) Furthermore, diagnostic practice differs according to location and changes over time, potentially influencing case identification.\(^5\) \(^2\) However, the recently completed UK Biobank mental health questionnaire is likely to provide much more robust data regarding mental disorders of all types.\(^5\) \(^3\)

**Disease surveillance**

Information gathered from electronic health records and other digital sources (internet searches, social media, and mobile phone data) has a huge potential for surveillance of mental disorders and their treatment.\(^5\) \(^4\) This could support the planning of services, implementation of interventions, evaluation of treatments, priority setting and the development of health policy and practice.

Traditional surveillance systems are often used in specific populations for specified exposures. They can be expensive to run, and data can be difficult to disseminate in a timely manner. However, real time acquisition and analysis of population, local and individual level data is possible with big data streams. This has the potential to improve
timeliness, resolution and access to hidden populations, providing a surveillance system for mental health previously not attainable. However there are significant challenges in using these data for surveillance in terms of the population sampled, their denominators, plus stage and severity of illness. Careful consideration would be required to align the aim of the surveillance system (screening, early detection, secondary prevention) with existing data sources, their completeness and quality. Given the challenges with routine coding of disease in mental health systems, integration of big data with validated survey output may be the way forward to improve timeliness but retain accuracy. This would also allow the inclusion of people not diagnosed or in contact with services. There is also a potential for surveillance of prescribing behaviours and adverse drug reactions. It would be possible to use the patient reporting of adverse reactions online (for example https://yellowcard.mhra.gov.uk and https://www.drugs.com), although checks and balances would be needed to prevent unfounded claims of adverse reactions impacting uptake of beneficial interventions.

DETECTION, SCREENING AND DIAGNOSIS

Diagnostic classification

Psychiatric disorders are traditionally classified into syndromes defined by expert consensus. An ideal diagnostic scheme would have consistency across settings and over time, and point to aetiology, prognosis and treatment response; current schemes do not achieve this, so a recorded diagnosis alone may be inadequate. Using full-text medical records for research with natural language processing can identify specific signs, symptoms, and health trajectories at a large scale. Studying these data could lead to better phenotypic classifications which predict clinically relevant outcomes. For example, the depression can be heterogeneous in prognosis: one study identified five broad trajectories of depression in 3000 patients using electronic health records.
Patients subsequently presenting with depression were sub-classified using the features identified to facilitate follow-up decisions.\textsuperscript{66}

However, clinical observations are often subjective. Therefore, there is also an effort towards both collecting and using objective and measurable data, such as neuroimaging and psychometrics for classification. The National Institute of Mental Health (NIMH) Research Domain Criteria programme (RDoC) encourages interdisciplinary study of psychopathological constructs postulated as relevant for the understanding of the mechanisms of mental disorders across categorical divides at the level of genes, cells, and circuits.\textsuperscript{17,67-69} Techniques from data science are needed to understand the resulting complexity. For example, a study taking an RDoC approach used behavioural, physiological and MRI measures in children, some with a clinical diagnosis of ADHD, and found three novel pathological phenotypes related to ADHD which cut across existing classifications: mild, extremely responsive to reward, and irritable, which could also be distinguished by patterns of cardiac reactivity and brain connectivity.\textsuperscript{70,71} Sources of information for could be extended to non-clinical domains such as social media and wearables to capture, for example sleep, physical activity, and shopping habits.\textsuperscript{72,73}

**Screening, detection and diagnosis**

An algorithm using coded fields and free text in electronic medical records can predict depression up to six months before the appearance of a coded diagnosis,\textsuperscript{74} representing an opportunity for automated screening. Since mental health disorders such as depression are often undiagnosed,\textsuperscript{75} screening would represent an opportunity for early identification and intervention, potentially reducing morbidity, saving lives, and providing economic benefits.\textsuperscript{76-78} However, all screening generates false positives and there is already concern about medicalising normality – such as feelings of stress or sadness – which could
undermine a person’s ability to cope, label them inappropriately, and result in unnecessary treatment.\textsuperscript{79-81} Thus, any screening model must be built upon mental health classifications that have the ability to distinguish if and when cases are likely to benefit from intervention.\textsuperscript{64,82} We hypothesise that such a model would need to use longitudinal clinical assessments and social context, alongside physiological, genetic and imaging data where available.

Such screening is some way from implementation, but a risk score for developing severe mental illness, or clinically relevant outcomes such as suicide, derived using data science could be possible within a few years.\textsuperscript{61,83-85} A data-driven approach has been taken to produce a risk score for cardiovascular events called QRISK2, using data from 2.3 million primary care records in the UK to produce an algorithm that uses parameters such as blood pressure, cholesterol and smoking status to estimate the likelihood of suffering a heart attack or stroke in the next decade.\textsuperscript{86} The QRISK2 score is now commonly presented to GPs when individual electronic health records are opened, enabling GPs to discuss how to reduce their risk. A mental health risk score could conceivably be similarly used, especially to flag particularly vulnerable people in high risk populations where mental health screening is already accepted, for example new mothers post-partum or people presenting with self-harm.\textsuperscript{87-89}

Outside the clinical context, wider phenotyping using non-medical data might also have potential for improving health outcomes, for example early detection of dementia or mild cognitive impairment via remote monitoring of patterns of behaviour (e.g. of social media posts, phone calls etc.)\textsuperscript{90} or mass screening of twitter posts for signs that someone is at risk of suicide.\textsuperscript{91} The advantages of using social media as a basis for screening, in younger people especially, is that they provide a setting to reach many millions of people
of diverse backgrounds. However, using people’s data without explicit consent needs careful consideration, as users of social media may not be comfortable for even “public” posts to be analysed in this way, and report being worried about potential stigmatisation, highlighting some of the potential ethical challenges of passive big data screening.

TREATMENT & THERAPY DEVELOPMENT

Figure 3 shows a four-stage cycle describing the application of data science in the treatment of mental illness, discussed in the following sections.

Targeted recruitment: Electronic health record systems as screening tools.

Patient recruitment is a rate-limiting step in clinical trials and one of the strongest drivers of costs. Consequently, many trials do not achieve recruitment targets. Harnessing the potential of electronic health records could expedite patient recruitment in mental health research: using routinely collected data as a screening patients for eligibility. In the South London and Maudsley NHS Foundation Trust’s “Consent for Contact” platform patients of this mental health trust are routinely asked for consent to be approached about relevant research projects based on information in their health record. These records can therefore be used to target and approach pre-consenting patients for potential studies. This model was evaluated for 2,106 participants, of whom 74.1% gave consent for contact. Furthermore, approaches identifying treatment-resistant groups – who are of growing interest in mental health research – could be done better in such large databases than clinicians alone.

Repurposing
Linking routinely collected health and administrative data to research data may provide new opportunities for treatment evaluations and repurposing of existing therapies for new indications. Longitudinal studies of individuals before and after receipt of an intervention may highlight unforeseen or off-target effects on mood or daily function that suggest efficacy beyond a drug's original indication. Whilst observational data are vulnerable to confounding, novel methods now exist to reduce this influence. For example, Mendelian Randomization can also be applied where there is linkage to genetic information.\textsuperscript{101} This approach has been shown to be useful, potentially preventing multi-million pound trials of interventions subsequently shown to be ineffective.\textsuperscript{102}

Further examples of how data science may lead to the repurposing of existing treatments for new mental health indications come again from genetics. Genetic studies of neuroticism and depression have shown enrichment of known genetic associations in the downstream targets of antidepressant drugs,\textsuperscript{103,104} providing an important ‘proof of concept’. Since the effects of currently effective treatments are enriched in the genetic associations of mental disorders, then genetic studies may also be able to identify new treatments and repurpose old ones. Genetic studies, such as GENDEP\textsuperscript{105} have also used treatment response or side effects as their phenotype of interest. These studies promise to reveal why some people respond better than others to treatment. They also provide the prospect of identifying who will respond best to a treatment, with fewest side effects, before prescribing.

**TREATMENT EVALUATION**

As well as streamlining trial execution, data science can directly evaluate the intervention. Observational epidemiology has contributed greatly to healthcare developments where trials were impossible, even before the age of data science, for example how babies
should be put in their cots to sleep and the health risks of smoking tobacco. With electronic case records, all patients generate data which can be used in vast observational studies, to allow for inexpensive and rapid improvements to health and healthcare, under suitable governance arrangements.

For example, cholinesterase inhibitors temporarily slow cognitive decline in people with Alzheimer’s dementia in randomised, controlled trials, but uncertainty remained about effects in real-world patients with multiple comorbidities. A ground-breaking study using pseudonymised healthcare records extracted text descriptions of cognitive test scores for 2460 patients prescribed cholinesterase inhibitors, and found similar treatment effects to the trials.\textsuperscript{106} This is a proof of the potential value of observational data, given large sample sizes, improved generalisability, and more complete follow up. This is particularly relevant in mental health research where participant disengagement may be greater.

Mental health treatment evaluation using routine data has been hindered by limited high-quality data on relevant treatment outcomes.\textsuperscript{107,108} Practice research networks bridge the gap between service provision and research and are uniquely placed to promote collection of quality healthcare data at scale in a way that is acceptable to patients, clinicians, and researchers. For example, the Child Outcomes Research Consortium is a collaboration of child mental health providers in the UK which collects and shares data with focused on patient reported measures of outcome and experience.\textsuperscript{109,110} The Northern Improving Access to Psychological Therapies (IAPT) Practice Research Network is a collaboration of psychotherapy service providers and research institutes promoting the use of data for service provision and research.\textsuperscript{111,112} Importantly, all IAPT services in the UK, which treat over 500,000 patients each year, collect common outcome measures at each session, consisting of short patient-completed questionnaires, providing data on therapy
progression. IAPT data have already delivered insights on the utility of a clinician-support system that alerts therapists to patients who are not responding as well as expected, and ongoing work is using patient characteristics to guide treatment choice.\textsuperscript{113,114} A further example is the U.S. Mental Health Research Network which brings together 13 health-system research centres providing care for 12.5 million people; early findings include behavioural activation being effective in perinatal depression\textsuperscript{115} and identifying subgroups less likely to adhere to antidepressant treatment.\textsuperscript{116}

Guidance published by organisations such as National Institute for Health and Care Excellence (NICE; England and Wales), Scottish Medicines Consortium, or Institut für Qualitaet und Wirtschaftlichkeit im Gesundheitswesen (IQWiG; Germany) raise standards, reduce variability, and provide a basis for monitoring. These organisations may currently value interventional research above observational research but new observational research – with the progress in ease (including regulatory reform),\textsuperscript{117} magnitude, speed, and methods suggesting causality facilitated by data science – will contribute more to the evidence about many healthcare interventions.\textsuperscript{118}

**DISEASE MANAGEMENT**

Three-quarters of the UK population own a smartphone (Ofcom) and growing numbers use wearable health devices/apps. New advances in technology are promising to transfer aspects of support and care from clinicians to patients. There is great potential to engage patients in their treatment and to move from sporadic patient contact towards continuous monitoring and guidance. In mental health, advances have been made in technology-assisted self-reporting and automated sensing.

**Self-reporting and Management**
Smartphone apps make remote patient-directed assessment of symptoms\textsuperscript{119} or other self-reported measures possible.\textsuperscript{120} Through data science techniques, some apps may provide a platform for intelligent assessment and recommendations tailored to the individual patient. However, the quality of and evidence base for health-related apps is variable. They can achieve quick assessment with real-time feedback and can establish a communication channel with carers and physicians as well as automated remote support. There is evidence that apps could reduce substance abuse, depression and stress.\textsuperscript{121}

There are a number of self-management apps available which enable self-tracking of mood to facilitate treatment or support patients to manage panic attacks.\textsuperscript{122}

**Passive Sensing & Analytics**

Emerging wearable sensor technologies offer real-time monitoring through continuously-collected data without patients having to do anything using the sensors in a smartphone/wearable device. For example, GPS traces can assess amount of time spent outdoors, accelerometers provide an indication of physical activity, and detection of other Bluetooth devices can estimate a person’s social contacts. Despite the field’s infancy, there exist a few applications of mobile/wearable devices to mental health: stress monitoring in everyday life and the workplace\textsuperscript{123,124}, early detection of Parkinson’s disease\textsuperscript{125}, and remote monitoring of sleep-awake activities to predict relapse in psychosis.\textsuperscript{126,127}

**HEALTH AND SOCIAL CARE RESEARCH**

The complexity and scale of health and social care means the use of information for service planning, delivery, and monitoring of outcomes is ever more crucial. Examples include understanding the impact of proposed changes to services, planning for changing population needs across services, and monitoring quality, safety, and equity of care within
and across sectors. For data to be meaningful, it must be possible to link them over time across different parts of the care pathway which may include both health and social care systems. This is particularly important in mental health where conditions are often long-term with health and social care needs. For example, understanding the relationship between particular conditions and service use can identify opportunities to prevent unscheduled care or identify inefficiencies across systems.

Mental health can be seen as being "constrained or facilitated by the social structures in which [a person is] positioned."¹²⁸ Five dimensions representing potential challenges to optimal social functioning have been identified: social integration; social contribution; social coherence; social actualization; and social acceptance.¹²⁸ This view of mental health shifts the focus from the individual to include the community and social structures within which people are located. Given this rich conceptualisation, other data are also necessary to understand mental health, including Census, environmental, housing, education, work and pensions, and crime data. Resources such as the Urban Big Data Centre (http://ubdc.ac.uk/) and the Administrative Data Research Network (https://www.adrn.ac.uk/) provide researchers with access to de-identified administrative data linked with health and social care data in a secure environment, such as that provided by Health Data Research UK and the Secure Anonymised Information Linkage (SAIL) Databank. Reflecting the complex arrangements for health and social care integration, agreement to share and link data across sectors is not always simple but efforts to secure such linkages have much to contribute to improving health and social care research and the social dimensions of health and health outcomes. However, as with other potential applications, there is a potential risk that solutions are driven by which data are available, resulting in policy makers getting a blinkered view of these processes. A better approach would be to ascertain what data are needed and then try to collect them.
ETHICAL CONSIDERATIONS
As discussed in many of the preceding sections, conducting mental health research using data science techniques brings ethical concerns surrounding privacy. However, we do not want mental health research to miss out on the breadth of opportunity outlined in this article. There are currently many checks and balances (including data protection legislation) in accessing any personal data, with additional rigorous processes in place for accessing data held by statutory bodies. The UK Biobank model of consent is an example of public willingness to consent to multiple uses of their data for research purposes. Others have proposed even broader ‘social contracts’ to enable data usage for public benefit. Whilst acknowledging some of the concerns as noted above, research demonstrates that the public can look favourably upon the use of social media data for health research, even for mental health, provided anonymity is ensured. Academic researchers may not be able to plan for, or resolve, all the potential ethical issues which the use of data science for mental health research may uncover, but it is without doubt that we must attempt to do this in consultation with, and support from, those living with mental illness. Indeed, we are extremely grateful for the contributions of people with lived experience of mental health problems in drafting this article.

CONCLUSION
Data science is a rapidly evolving field which offers many valuable applications to mental health research, examples of which we have outlined in this article. Most importantly, it offers the possibility of making research incorporating real-world complexity tractable. We anticipate that the substantial advancements in mental health research we are beginning to see will bring tangible benefits to people with mental illness.
REFERENCES


Insel, T. R. The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. - PubMed - NCBI.


Knapp, M., McDaid, D. & Parsonage, M. *Mental health promotion and mental illness prevention: the economic case*. (London School of Economics and Political Science; Centre for Mental Health; Centre for the Economics of Mental Health, Institute of Psychiatry, King's College London, 2011).


Taggart, H. THE FIVE YEAR FORWARD VIEW FOR MENTAL HEALTH. (2016).


Ibrahim, Z. M. et al. A multi-agent platform for automating the collection of patient-provided clinical feedback in *Proceedings of the 2015 International Conference on Autonomous Agents and


BOX 1. UKCRC Health Research Classification System Research Activity Codes
(http://www.hrcsonline.net/rac)

1. **Underpinning**: Research that underpins investigations into the cause, development, detection, treatment and management of diseases, conditions and ill health

2. **Aetiology**: Identification of determinants that are involved in the cause, risk or development of disease, conditions and ill health

3. **Prevention**: Research aimed at the primary prevention of disease, conditions or ill health, or promotion of well-being

4. **Detection & Diagnosis**: Discovery, development and evaluation of diagnostic, prognostic and predictive markers and technologies

5. **Treatment Development**: Discovery and development of therapeutic interventions and testing in model systems and preclinical settings

6. **Treatment Evaluation**: Testing and evaluation of therapeutic interventions in clinical, community or applied settings

7. **Disease Management**: Research into individual care needs and management of disease, conditions or ill health

8. **Health Services**: Research into the provision and delivery of health and social care services, health policy and studies of research design, measurements and methodologies
FIGURE 1. What can data science do for mental health research? Data science is key to improving diagnosis, transforming treatments and ultimately making mental illness preventable.
FIGURE 2. Data science applications in understanding mental health and mental illness – understanding biopsychosocial processes in health informs our understanding of aetiology which leads on to prevention initiatives which require robust mechanisms for detection.
FIGURE 3. The cycle of data science applications in the context of mental health treatments – more efficient and targeted recruitment to trials leads on to large-scale evaluation of treatments and assessment of real-world efficacy which could highlight opportunities for drug repurposing.
AUTHOR CONTRIBUTIONS: All authors drafted individual sections of the manuscript and revised it in its entirety for final content.

ACKNOWLEDGMENTS: The MQ Data Science group was set up by the UK mental health research charity MQ in 2015 and includes UK-based researchers from a range of disciplines working the field of mental health data science. The authors of this article are all members of the MQ Data Science group and the article stemmed from discussions at a previous meeting of the wider group.

TCR is a member of the Alzheimer Scotland Dementia Research Centre funded by Alzheimer Scotland. TCR and AMM are both members of the University of Edinburgh Centre for Cognitive Ageing & Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (G0700704/84698). Funding from the Biotechnology and Biological Sciences Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, and Medical Research Council is gratefully acknowledged for the latter. KASD, ZI, and RS are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. WL is supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula (NIHR CLAHRC South West Peninsula). AMM has received funding from the Sackler Trust, the Wellcome Trust, and an MRC Mental Health Data Pathfinder award (MC_PC_17209). RS has received research funding in the last 3 years from Janssen, Roche and GSK.

The views expressed are those of the authors and not necessarily those of the NHS, the National Institute of Health Research, the Department of Health and Social Care, or any other funder.
**ROLE OF FUNDER:** MQ sponsored the meetings from which this paper emerged. Other than one of the authors (EW) being employed by MQ, the charity had no role in the preparation of the manuscript and the final decision to publish was made by the corresponding author.

**COMPETING INTERESTS:** Sources of funding are mentioned above. The authors declare no other competing interests.