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 John [7], William Lee,[12,13] Margaret Maxwell,[14] Andrew M. McIntosh,[1,2] Rob Stewart,[7,8] and the MQ Data Science group[6]

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
8. South London and Maudsley NHS Foundation Trust, 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.[4] 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),[5] for example, has identified 148 independent genetic loci associated with variations in cognitive function,[6] which has relevance to researchers trying to understand how cognitive processes can be impaired in mental disorders.[7] Other examples include the US Million Veteran's Program (current N=600,000)[8] and the BioBank Japan project (N=200,000).[9] 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,[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[11]. Combined with the related discipline of bioinformatics, data science offers great potential for underpinning research through *in silico* discovery science, such as machine-learning-driven modelling of cellular protein folding[12] and synaptic transmission,[13] or computational modelling of receptor pharmacology for drug development.[14] In addition to further stimulating research on the molecular biology of psychiatric disorders, this work also supplements preclinical psychopharmacological research.[15]

Similarly, progress in neuroimaging has been constrained by a lack of sufficiently powered and consistent datasets of healthy controls[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][17], UK Biobank [Target N=100,000][5], and the Human Connectome Project[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.[19] Alteration in the default-mode network has been implicated in follow-up research in autism,[20] schizophrenia[21] and Alzheimer's dementia[22] among others.[23] Looking to the future, combining multiband imaging with genomics to study the inter-relationships between genotype, brain structure and functional architecture[24] is a petabyte-scale data science challenge which can now be undertaken.[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,[26] vital signs variation,[27] and alcohol use[28]. As well as guiding research in areas such as circadian rhythms in mental illness,[29] such tools can also be used to augment clinical care.[30] Diverse phenotypic data can also be extracted through data mining techniques deployed to social media and health records databases.[31,32] Finally, record linkage to repositories of social and educational data[33] will enable construction of a truly integrative model of mental processes, in health as well as disease.[34]

AETIOLOGY, PREVENTION OF DISEASE AND PROMOTION OF WELL-BEING

All mental disorders have a complex aetiology.[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.[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.[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.[40] 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.[41] A number of vitamins, including folic acid and vitamins B_{12} and D have been linked with mental illnesses.[42,43] Mendelian randomization has also been used to infer a causal relationship of lower vitamin D with Alzheimer's disease[44] 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.[45] 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.[46-48] There might also be scope to identify educational and occupational status from electronic health and other records.[49] 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[50] 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.[51] Furthermore, diagnostic practice differs according to location and changes over time, potentially influencing case identification.[52] However, the recently completed UK Biobank mental health questionnaire is likely to provide much more robust data regarding mental disorders of all types.[53]

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.[54] 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.[55] 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.[56,57] 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,[58,59] so a recorded diagnosis alone may be inadequate.[60] Using full-text medical records for research with natural language processing can identify specific signs, symptoms, and health trajectories at a large scale.[32,61,62] Studying these data could lead to better phenotypic classifications which predict clinically relevant outcomes.[63,64] For example, the depression can be heterogeneous in prognosis: one study identified five broad trajectories of depression in 3000 patients using electronic health records .[65]

Patients subsequently presenting with depression were sub-classified using the features identified to facilitate follow-up decisions.[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.[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.[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.[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,[74] representing an opportunity for automated screening. Since mental health disorders such as depression are often undiagnosed,[75] screening would represent an opportunity for early identification and intervention, potentially reducing morbidity, saving lives, and providing economic benefits.[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.[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.[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.[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.[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.[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.)[90] or mass screening of twitter posts for signs that someone is at risk of suicide.[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.[92] 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,[93,94] 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.[95] Consequently, many trials do not achieve recruitment targets.[96] 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.[97] 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.[98] This model was evaluated for 2,106 participants, of whom 74.1% gave consent for contact. [99,100] 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 drugs 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.[101] This approach has been shown to be useful, potentially preventing multi-million pound trials of interventions subsequently shown to be ineffective.[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,[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[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.[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.[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.[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.[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.[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[115] and identifying subgroups less likely to adhere to antidepressant treatment.[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),[117] magnitude, speed, and methods suggesting causality facilitated by data science – will contribute more to the evidence about many healthcare interventions.[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[119] or other self-reported measures possible.[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.[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.[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[123,124], early detection of Parkinson's disease[125], and remote monitoring of sleep-awake activities to predict relapse in psychosis.[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[129,130] 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.[131-133] 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,[93,94] 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.[134,135] 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

1 Walesby, K. E., Harrison, J. K. & Russ, T. C. What big data could achieve in Scotland. *The journal of the Royal College of Physicians of Edinburgh* **47**, 114-119, doi:10.4997/jrcpe.2017.201 (2017).

2 Soni, J., Ansari, U., Sharma, D. & Soni, S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications* **17**, 43-48 (2011).

3 Hamada, T., Keum, N., Nishihara, R. & Ogino, S. Molecular pathological epidemiology: new developing frontiers of big data science to study etiologies and pathogenesis. *Journal of Gastroenterology* **52**, 265-275, doi:10.1007/s00535-016-1272-3 (2017).

4 Hafferty, J. D., Smith, D. J. & McIntosh, A. M. Invited Commentary on Stewart and Davis "'Big data'in mental health research—current status and emerging possibilities". *Social psychiatry and psychiatric epidemiology* **52**, 127-129 (2017).

5 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779 (2015).

6 Davies, G. *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature communications* **9**, 2098, doi:10.1038/s41467-018-04362-x (2018).

7 Lencz, T. *et al.* Molecular genetic evidence for overlap between general cognitive ability and risk for schizophrenia: a report from the Cognitive Genomics consorTium (COGENT). *Mol Psychiatry* **19**, 168-174, doi:10.1038/mp.2013.166 (2014).

8 Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology* **70**, 214-223, doi:10.1016/j.jclinepi.2015.09.016 (2016).

9 Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *Journal of epidemiology* **27**, S2-s8, doi:10.1016/j.je.2016.12.005 (2017).

10 Herper, M. *Illumina Promises To Sequence Human Genome For $100 - But Not Quite Yet*, <https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/#13a4d50d386d> (2017).

11 Schatz, M. C. Biological data sciences in genome research. *Genome research* **25**, 1417-1422 (2015).

12 Cheng, J., Tegge, A. N. & Baldi, P. Machine learning methods for protein structure prediction. *IEEE reviews in biomedical engineering* **1**, 41-49 (2008).

13 Montes, J., Gomez, E., Merchán-Pérez, A., DeFelipe, J. & Peña, J.-M. A machine learning method for the prediction of receptor activation in the simulation of synapses. *PloS one* **8**, e68888 (2013).

14 Ou-Yang, S.-s. *et al.* Computational drug discovery. *Acta Pharmacologica Sinica* **33**, 1131 (2012).

15 Immadisetty, K., Geffert, L. M., Surratt, C. K. & Madura, J. D. New design strategies for antidepressant drugs. *Expert opinion on drug discovery* **8**, 1399-1414, doi:10.1517/17460441.2013.830102 (2013).

16 Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience* **19**, 1523-1536 (2016).

17 Thompson, P. M. *et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior* **8**, 153-182 (2014).

18 Shi, Y. & Toga, A. Connectome imaging for mapping human brain pathways. *Molecular Psychiatry* (2017).

19 Gu, S. *et al.* Emergence of system roles in normative neurodevelopment. *Proceedings of the National Academy of Sciences* **112**, 13681-13686 (2015).

20 Christakou, A. *et al.* Disorder-specific functional abnormalities during sustained attention in youth with Attention Deficit Hyperactivity Disorder (ADHD) and with autism. *Mol Psychiatry* **18**, 236-244, doi:10.1038/mp.2011.185 (2013).

21 Chang, X. *et al.* Altered default mode and fronto-parietal network subsystems in patients with schizophrenia and their unaffected siblings. *Brain research* **1562**, 87-99, doi:10.1016/j.brainres.2014.03.024 (2014).

22 Schwindt, G. C. *et al.* Modulation of the default-mode network between rest and task in Alzheimer's Disease. *Cerebral cortex (New York, N.Y. : 1991)* **23**, 1685-1694, doi:10.1093/cercor/bhs160 (2013).

23 Broyd, S. J. *et al.* Default-mode brain dysfunction in mental disorders: a systematic review. *Neuroscience and biobehavioral reviews* **33**, 279-296, doi:10.1016/j.neubiorev.2008.09.002 (2009).

24 Xia, M. & He, Y. Functional connectomics from a "big data" perspective. *NeuroImage* (2017).

25 Van Horn, J. D. & Toga, A. W. Human neuroimaging as a "Big Data" science. *Brain imaging and behavior* **8**, 323-331 (2014).

26 Bidargaddi, N. *et al.* Digital footprints: facilitating large-scale environmental psychiatric research in naturalistic settings through data from everyday technologies. *Molecular psychiatry* **22**, 164 (2017).
27 Khan, Y., Ostfeld, A. E., Lochner, C. M., Pierre, A. & Arias, A. C. Monitoring of vital signs with flexible and wearable medical devices. *Advanced Materials* **28**, 4373-4395 (2016).
28 Selvam, A. P., Muthukumar, S., Kamakoti, V. & Prasad, S. A wearable biochemical sensor for monitoring alcohol consumption lifestyle through Ethyl glucuronide (EtG) detection in human sweat. *Scientific reports* **6**, 23111 (2016).
29 Bradley, A. J. *et al.* Sleep and circadian rhythm disturbance in bipolar disorder. *Psychol Med* **47**, 1678-1689, doi:10.1017/s0033291717000186 (2017).
30 Knight, A. & Bidargaddi, N. Commonly available activity tracker apps and wearables as a mental health outcome indicator: A prospective observational cohort study among young adults with psychological distress. *J Affect Disord* **236**, 31-36, doi:10.1016/j.jad.2018.04.099 (2018).
31 Zafarani, R., Abbasi, M. A. & Liu, H. *Social media mining: an introduction*. (Cambridge University Press, 2014).
32 Jackson, R. G. *et al.* Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* **7**, e012012 (2017).
33 Ford, D. V. *et al.* The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC health services research* **9**, 157 (2009).
34 McIntosh, A. M. *et al.* Data science for mental health: a UK perspective on a global challenge. *The lancet. Psychiatry* **3**, 993-998, doi:10.1016/s2215-0366(16)30089-x (2016).
35 Engel, G. The need for a new medical model: a challenge for biomedicine. *Science* **196**, 129-136, doi:10.1126/science.847460 (1977).
36 Astell-Burt, T., Mitchell, R. & Hartig, T. The association between green space and mental health varies across the lifecourse. A longitudinal study. *J Epidemiol Community Health* **68**, 578-583 (2014).
37 Gascon, M. *et al.* Mental health benefits of long-term exposure to residential green and blue spaces: a systematic review. *International journal of environmental research and public health* **12**, 4354-4379 (2015).
38 Ruijsbroek, A. *et al.* Neighbourhood green space, social environment and mental health: an examination in four European cities. *International journal of public health*, 1-11 (2017).
39 Chen, J. C. *et al.* Ambient air pollution and neurotoxicity on brain structure: evidence from women's health initiative memory study. *Annals of neurology* **78**, 466-476 (2015).
40 Miller, H. J. & Tolle, K. Big data for healthy cities: Using location-aware technologies, open data and 3D urban models to design healthier built environments. *Built Environment* **42**, 441-456 (2016).
41 Inoue, T. *et al.* Does temperature or sunshine mediate the effect of latitude on affective temperaments? A study of 5 regions in Japan. *Journal of affective disorders* **172**, 141-145 (2015).
42 Roffman, J. L., Lamberti, J., Achtyes, E. & et al. Randomized multicenter investigation of folate plus vitamin b12 supplementation in schizophrenia. *JAMA Psychiatry* **70**, 481-489, doi:10.1001/jamapsychiatry.2013.900 (2013).
43 Milaneschi, Y. *et al.* The association between low vitamin D and depressive disorders. *Molecular psychiatry* **19**, 444 (2014).
44 Mokry, L. E. *et al.* Genetically decreased vitamin D and risk of Alzheimer disease. *Neurology* **87**, 2567-2574 (2016).
45 Chen, H. *et al.* Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study. *The Lancet* **389**, 718-726 (2017).
46 White, J. *et al.* Improving Mental Health Through the Regeneration of Deprived Neighborhoods: A Natural Experiment. *American Journal of Epidemiology* **186**, 473-480, doi:10.1093/aje/kwx086 (2017).
47 Pettit, S. *et al.* Variation in referral and access to new psychological therapy services by age: an empirical quantitative study. *British Journal of General Practice* **67**, e453-e459, doi:10.3399/bjgp17X691361 (2017).
48 Asthana, S. *et al.* Equity of utilisation of cardiovascular care and mental health services in England: a cohort-based cross-sectional study using small-area estimation. (2016).
49 Wu, C.-Y. *et al.* Evaluation of Smoking Status Identification Using Electronic Health Records and Open-Text Information in a Large Mental Health Case Register. *PLOS ONE* **8**, e74262, doi:10.1371/journal.pone.0074262 (2013).

50 Vuorilehto, M. S., Melartin, T. K., Rytsala, H. J. & Isometsa, E. T. Do characteristics of patients with major depressive disorder differ between primary and psychiatric care? *Psychol Med* **37**, 893-904, doi:10.1017/s0033291707000098 (2007).
51 Demyttenaere, K. *et al.* Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Jama* **291**, 2581-2590, doi:10.1001/jama.291.21.2581 (2004).
52 John, A. *et al.* Case-finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. *BMC medical informatics and decision making* **16**, 35, doi:10.1186/s12911-016-0274-7 (2016).
53 Davis, K. A. S. *et al.* Mental Health in UK Biobank – implementation and results of an online questionnaire in 157,366 participants. *Brit J Psychiatr Open* (In Press).
54 Gregory E. Simon *et al.* First Presentation With Psychotic Symptoms in a Population-Based Sample. *Psychiatric Services* **68**, 456-461, doi:10.1176/appi.ps.201600257 (2017).
55 John, A. *et al.* Recent trends in primary-care antidepressant prescribing to children and young people: an e-cohort study. *Psychological Medicine* **46**, 3315-3327, doi:10.1017/S0033291716002099 (2016).
56 Aragona, M. Rethinking received views on the history of psychiatric nosology: minor shifts, major continuities. *Alternative perspectives on psychiatric validation. DSM, ICD, RDoC, and beyond*, 27-46 (2014).
57 Reed, G. M. *et al.* The ICD-11 developmental field study of reliability of diagnoses of high-burden mental disorders: results among adult patients in mental health settings of 13 countries. *World psychiatry* **17**, 174-186 (2018).
58 Blumenthal-Barby, J. Psychiatry's new manual (DSM-5): ethical and conceptual dimensions. *Journal of medical ethics*, medethics-2013-101468 (2013).
59 Ghaemi, S. N. Nosologomania: DSM & Karl Jaspers' Critique of Kraepelin. *Philosophy, Ethics, and Humanities in Medicine* **4**, 10 (2009).
60 Davis, K. A., Sudlow, C. L. & Hotopf, M. Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC psychiatry* **16**, 263 (2016).
61 National Information Board & Department of Health. *Personalised health and care 2020: Using Data and Technology to Transform Outcomes for Patients and Citizens*. (GOV.UK and NHS, HM Government, 2014).
62 Spiranovic, C., Matthews, A., Scanlan, J. & Kirkby, K. C. Increasing knowledge of mental illness through secondary research of electronic health records: opportunities and challenges. *Advances in Mental Health* **14**, 14-25, doi:10.1080/18387357.2015.1063635 (2016).
63 Frances, A. J. & Widiger, T. Psychiatric diagnosis: lessons from the DSM-IV past and cautions for the DSM-5 future. *Annual review of clinical psychology* **8**, 109-130 (2012).
64 Hickie, I. B. *et al.* Clinical classification in mental health at the cross-roads: which direction next? *BMC medicine* **11**, 125 (2013).
65 Lin, Y., Huang, S., Simon, G. E. & Liu, S. Analysis of depression trajectory patterns using collaborative learning. *Mathematical Biosciences* **282**, 191-203, doi:https://doi.org/10.1016/j.mbs.2016.10.008 (2016).
66 Lin, Y., Huang, S., Simon, G. E. & Liu, S. Data-based Decision Rules to Personalize Depression Follow-up. *Scientific reports* **8**, 5064 (2018).
67 Insel, T. R. The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. - PubMed - NCBI.
68 Deisseroth, K. Circuit dynamics of adaptive and maladaptive behaviour. *Nature* **505**, 309 (2014).
69 KCL Institute of Psychiatry Psychology & Neuroscience. *Psychometrics & Measurement Database*, <http://www.kcl.ac.uk/ioppn/depts/BiostatisticsHealthInformatics/Psychometrics-and-measurement-lab/Psychometrics-and-measurement-database.aspx> (2017).
70 Carcone, D. & Ruocco, A. C. Six years of research on the National Institute of Mental Health's Research domain criteria (RDoC) initiative: A systematic review. *Frontiers in cellular neuroscience* **11**, 46 (2017).
71 Karalunas, S. L. *et al.* Subtyping attention-deficit/hyperactivity disorder using temperament dimensions: toward biologically based nosologic criteria. *JAMA psychiatry* **71**, 1015-1024 (2014).
72 Casey, J. A., Schwartz, B. S., Stewart, W. F. & Adler, N. E. Using electronic health records for population health research: a review of methods and applications. *Annual review of public health* **37**, 61-81 (2016).

73 Torous, J., Onnela, J. & Keshavan, M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Translational psychiatry* **7**, e1053 (2017).
74 Huang, S. H. *et al.* Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association* **21**, 1069-1075 (2014).
75 Whiteford, H. A. *et al.* Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet* **382**, 1575-1586 (2013).
76 Friedman, R. A. Uncovering an epidemic—screening for mental illness in teens. *New England Journal of Medicine* **355**, 2717-2719 (2006).
77 Hetrick, S. *et al.* Early identification and intervention in depressive disorders: towards a clinical staging model. *Psychotherapy and psychosomatics* **77**, 263-270 (2008).
78 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).
79 Parker, G. Head to head: Is depression overdiagnosed? Yes. *BMJ: British Medical Journal* **335**, 328 (2007).
80 Najman, J. M. *et al.* Screening in early childhood for risk of later mental health problems: A longitudinal study. *Journal of psychiatric research* **42**, 694-700 (2008).
81 Henderson, S. W., Horwitz, A. V. & Wakefield, J. C. Should screening for depression among children and adolescents be demedicalized? *Journal of the American Academy of Child & Adolescent Psychiatry* **48**, 683-687 (2009).
82 McGorry, P. D. Staging in neuropsychiatry: a heuristic model for understanding, prevention and treatment. *Neurotoxicity research* **18**, 244-255 (2010).
83 Schoevers, R. A. *et al.* Prevention of late-life depression in primary care: do we know where to begin? *American Journal of Psychiatry* **163**, 1611-1621 (2006).
84 Nock, M. K. *et al.* Cross-national analysis of the associations among mental disorders and suicidal behavior: findings from the WHO World Mental Health Surveys. *PLoS medicine* **6**, e1000123 (2009).
85 Poulin, C. *et al.* Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one* **9**, e85733 (2014).
86 Hippisley-Cox, J. *et al.* Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* **336**, 1475-1482 (2008).
87 Olfson, M., Marcus, S. C. & Bridge, J. A. Emergency department recognition of mental disorders and short-term outcome of deliberate self-harm. *American Journal of Psychiatry* **170**, 1442-1450 (2013).
88 College of Emergency Medicine. *Mental Health in Emergency Departments: a toolkit for improving care*. (The College of Emergency Medicine, 2013).
89 National Institute for Health and Clinical Excellence. *Antenatal and Postnatal Mental Health: Clinical management and service guidance (update) CG192*. (National Institute for Health and Clinical Excellence, 2014).
90 Kaye, J. *et al.* Unobtrusive measurement of daily computer use to detect mild cognitive impairment. *Alzheimer's & Dementia* **10**, 10-17 (2014).
91 Jashinsky, J. *et al.* Tracking suicide risk factors through Twitter in the US. *Crisis: The Journal of Crisis Intervention and Suicide Prevention* **35**, 51 (2014).
92 Inkster, B., Stillwell, D., Kosinski, M. & Jones, P. A decade into Facebook: where is psychiatry in the digital age? *The Lancet Psychiatry* **3**, 1087-1090 (2016).
93 Conway, M. & O'Connor, D. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology* **9**, 77-82 (2016).
94 Mikal, J., Hurst, S. & Conway, M. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics* **17**, 22 (2016).
95 (US), I. o. M. *Recruitment Challenges in Clinical Trials for Different Diseases and Conditions*. (National Academies Press (US), 2012).
96 McDonald, A. M. *et al.* What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials* **7**, 9 (2006).
97 McGregor, J. *et al.* The Health Informatics Trial Enhancement Project (HITE): Using routinely collected primary care data to identify potential participants for a depression trial. *Trials* **11**, 39 (2010).
98 Callard, F. *et al.* Developing a new model for patient recruitment in mental health services: a cohort study using Electronic Health Records. *BMJ open* **4**, e005654 (2014).

99 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* **14**, 377 (2014).
100 Robotham, D. *et al.* Facilitating mental health research for patients, clinicians and researchers: a mixed-method study. *BMJ open* **6**, e011127 (2016).
101 Relton, C. L. & Davey Smith, G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International journal of epidemiology* **41**, 161-176, doi:10.1093/ije/dyr233 (2012).
102 Burgess, S., Butterworth, A., Malarstig, A. & Thompson, S. G. Use of Mendelian randomisation to assess potential benefit of clinical intervention. *BMJ : British Medical Journal* **345**, doi:10.1136/bmj.e7325 (2012).
103 Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668-681, doi:10.1038/s41588-018-0090-3 (2018).
104 Luciano, M. *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat Genet* **50**, 6-11, doi:10.1038/s41588-017-0013-8 (2018).
105 Fabbri, C. *et al.* New insights into the pharmacogenomics of antidepressant response from the GENDEP and STAR*D studies: rare variant analysis and high-density imputation. *The pharmacogenomics journal* **18**, 413-421, doi:10.1038/tpj.2017.44 (2018).
106 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* **9**, e109484 (2014).
107 Taggart, H. THE FIVE YEAR FORWARD VIEW FOR MENTAL HEALTH. (2016).
108 Maddox, T. M. & Ferguson, T. B. The Potential of Learning Health Care Systems: The SWEDEHEART Example∗. *Journal of the American College of Cardiology* **66**, 544-546, doi:https://doi.org/10.1016/j.jacc.2015.05.050 (2015).
109 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. *Administration and Policy in Mental Health and Mental Health Services Research* **43**, 297-301, doi:10.1007/s10488-014-0592-y (2016).
110 Clark, D. M. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry* **23**, 318-327, doi:10.3109/09540261.2011.606803 (2011).
111 Lucock, M. *et al.* The Role of Practice Research Networks (PRN) in the Development and Implementation of Evidence: The Northern Improving Access to Psychological Therapies PRN Case Study. *Administration and Policy in Mental Health and Mental Health Services Research* **44**, 919-931, doi:10.1007/s10488-017-0810-5 (2017).
112 Delgadillo, J. *et al.* Improving the efficiency of psychological treatment using outcome feedback technology. *Behaviour Research and Therapy* **99**, 89-97, doi:https://doi.org/10.1016/j.brat.2017.09.011 (2017).
113 DeRubeis, R. J. *et al.* The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *PLOS ONE* **9**, e83875, doi:10.1371/journal.pone.0083875 (2014).
114 Saunders, R., Cape, J., Fearon, P. & Pilling, S. Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *J Affect Disord* **197**, 107-115, doi:10.1016/j.jad.2016.03.011 (2016).
115 Dimidjian, S. *et al.* A pragmatic randomized clinical trial of behavioral activation for depressed pregnant women. *Journal of consulting and clinical psychology* **85**, 26 (2017).
116 Rossom, R. C. *et al.* ANTIDEPRESSANT ADHERENCE ACROSS DIVERSE POPULATIONS AND HEALTHCARE SETTINGS. *Depression and anxiety* **33**, 765-774, doi:10.1002/da.22532 (2016).
117 National Institute for Health and Clinical Excellence. *DATA SCIENCE FOR HEALTH AND CARE EXCELLENCE: Harnessing the UK opportunities for new research and decision-making paradigms*. (National Institute for Health and Clinical Excellence, 2016).
118 Gillan, C. M. & Whelan, R. What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences* **18**, 34-42 (2017).
119 Gravenhorst, F. *et al.* Mobile phones as medical devices in mental disorder treatment: an overview. *Personal and Ubiquitous Computing* **19**, 335-353 (2015).
120 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*

Multiagent Systems. 831-839 (International Foundation for Autonomous Agents and Multiagent Systems).
121 Donker, T. *et al.* Smartphones for Smarter Delivery of Mental Health Programs: A Systematic Review. *Journal of Medical Internet Research* **15**, e247, doi:10.2196/jmir.2791 (2013).
122 Marley, J. & Farooq, S. Mobile telephone apps in mental health practice: uses, opportunities and challenges. *BJPsych Bull* **39**, 288-290 (2015).
123 Muaremi, A., Arnrich, B. & Tröster, G. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* **3**, 172-183 (2013).
124 Muaremi, A., Bexheti, A., Gravenhorst, F., Arnrich, B. & Tröster, G. Monitoring the impact of stress on the sleep patterns of pilgrims using wearable sensors in *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on.* 185-188 (IEEE).
125 Mazilu, S. *et al.* GaitAssist: a daily-life support and training system for parkinson's disease patients with freezing of gait in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems.* 2531-2540 (ACM).
126 Meyer, N. *et al.* Detecting Early Signs of Relapse in Psychosis Using Remote Monitoring Technology: Acceptability and Feasibility of a Passive Sensing Approach. *Early Intervention in Psychiatry* **10**, 112-112 (2016).
127 Kerz, M. *et al.* SleepSight: A wearables-based relapse prevention system for Schizophrenia in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct.* 113-116 (ACM).
128 Keyes, J. *Banking technology handbook*. (CRC Press, 1998).
129 Laurie, G. *et al.* On moving targets and magic bullets: Can the UK lead the way with responsible data linkage for health research? *International journal of medical informatics* **84**, 933-940 (2015).
130 Nuffield Council on Bioethics. *The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues*. (Nuffield Council on Bioethics, 2015).
131 Hemingway, H. *et al.* Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *European Heart Journal* (2017).
132 Carter, P., Laurie, G. T. & Dixon-Woods, M. The social licence for research: why care. data ran into trouble. *Journal of medical ethics*, medethics-2014-102374 (2015).
133 Sethi, N. & Laurie, G. T. Delivering proportionate governance in the era of eHealth: making linkage and privacy work together. *Medical law international* **13**, 168-204 (2013).
134 Jones, K. H., McNerney, C. L. & Ford, D. V. Involving consumers in the work of a data linkage research unit. *International Journal of Consumer Studies* **38**, 45-51, doi:10.1111/ijcs.12062 (2014).
135 Ennis, L. & Wykes, T. Impact of patient involvement in mental health research: longitudinal study. *British Journal of Psychiatry* **203**, 381-386, doi:10.1192/bjp.bp.112.119818 (2018).

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

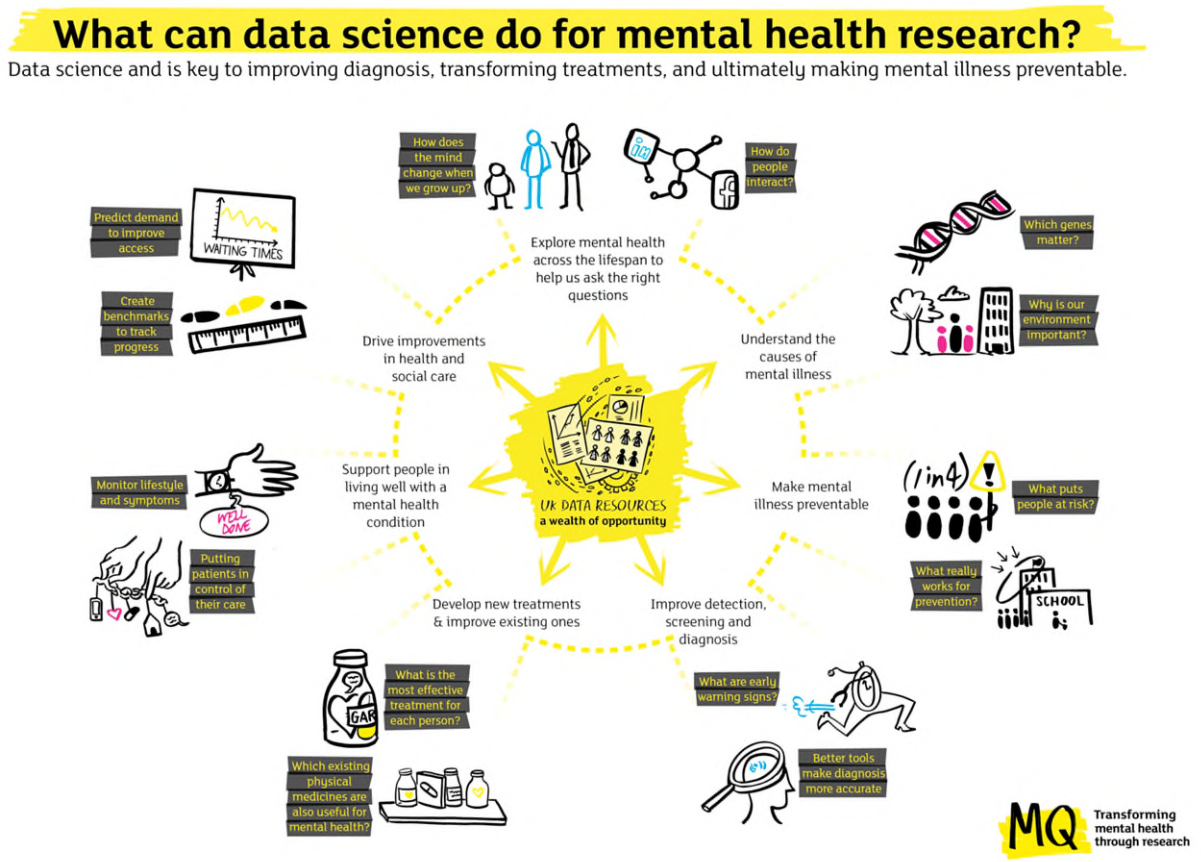


FIGURE 2. Data science applications in understanding mental health and mental illness – understanding biopsychosocial processes in health informs our understanding of aetiology which leads on to prevention initiatives which require robust mechanisms for detection

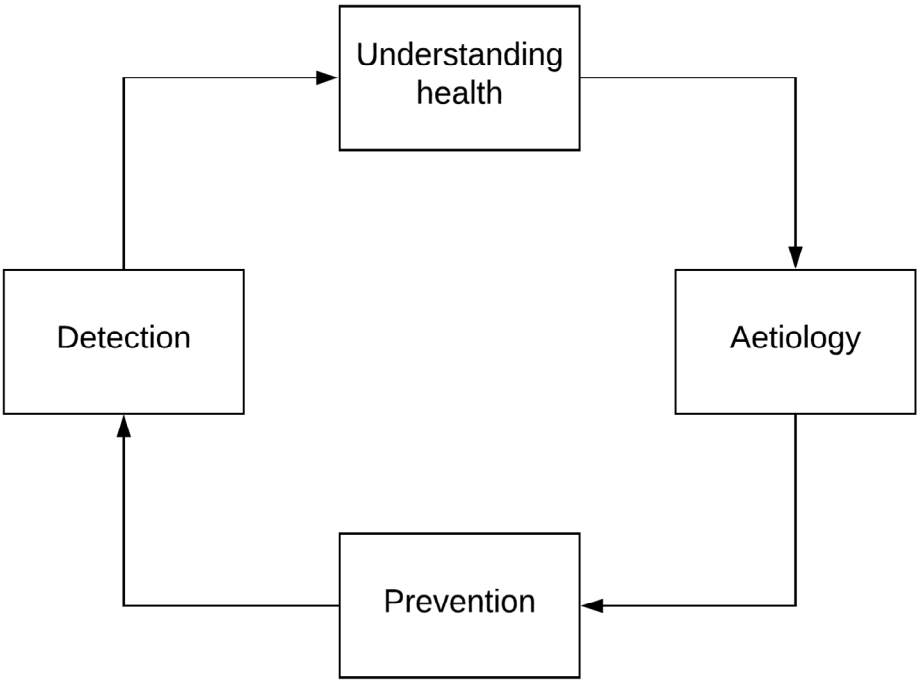
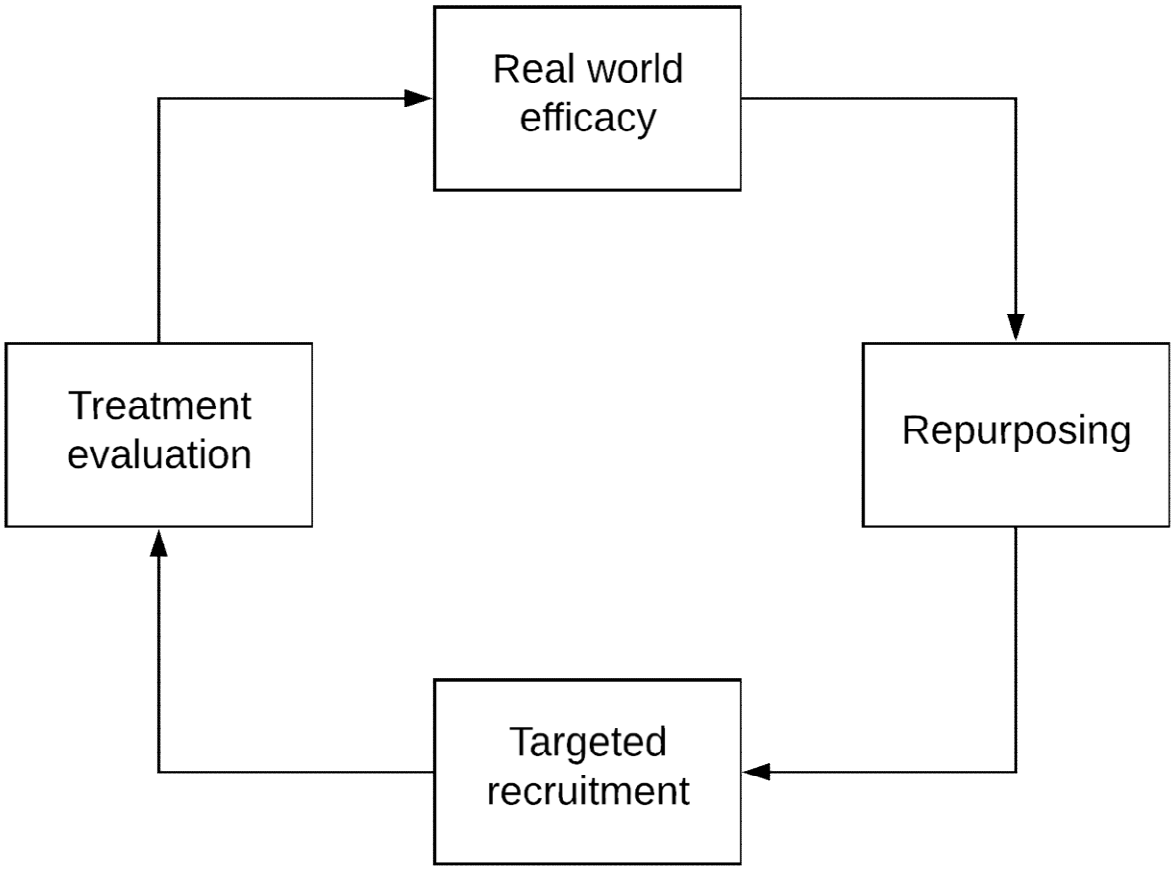


FIGURE 3. The cycle of data science applications in the context of mental health treatments – more efficient and targeted recruitment to trials leads on to large-scale evaluation of treatments and assessment of real-world efficacy which could highlight opportunities for drug repurposing



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.