BJPsych Advances (2018), vol. 24, 237–244 doi: 10.1192/bja.2018.15

Big data: what it can and cannot achieve

ARTICLE

Peter Schofield & Jayati Das-Munshi

SUMMARY

This article looks at the use of large datasets of health records, typically linked with other data sources, in mental health research. The most comprehensive examples of this kind of 'big data' are typically found in Scandinavian countries, although there are also many useful sources in the UK. There are a number of promising methodological innovations from studies using big data in UK mental health research, including: hybrid study designs, data linkage and enhanced study recruitment. It is, however, important to be aware of the limitations of research using big data, particularly the various pitfalls in analysis. We therefore caution against abandoning traditional research designs, and argue that other data sources are equally valuable and, ideally, research should incorporate data from a range of sources.

LEARNING OBJECTIVES

- Be aware of major big data resources relevant to mental health research
- Be aware of key advantages and innovative study designs using these data sources
- Understand the inherent limitations to studies reliant on big data alone

DECLARATION OF INTEREST

None.

In recent years much has been written about 'big data', to such an extent that the literature on this topic is now almost as dizzying in magnitude as the data that are written about. In this short article we aim to highlight in a non-technical way some of the advantages and disadvantages of these resources, both for those who are actively involved in research and for clinicians who need to assess the value and clinical relevance of research evidence. This article concentrates on one kind of big data: large datasets of health records, typically linked to other large datasets, including administrative and census data. This kind of data already plays an important part in psychiatric research and its role is expanding rapidly.

What do we mean by big data?

Arguably the most successful examples in psychiatric research have come from Scandinavian population registers. These comprise health records data collected for the entire population over many years, linked to a range of administrative data. They have a particular value for psychiatric research for a number of reasons: they provide information about those who would otherwise be hard to reach using conventional survey approaches, their scale makes it possible to answer questions about disorders that are relatively rare and, with data often collected over a long period, we can look at mental health outcomes independently of exposures. The last is particularly important when studying risk factors for severe mental illness. For example, for some time studies have shown elevated rates of psychosis in urban areas, although this could be simply the effect of 'social drift', where those who are ill or in the prodromal phase 'drift' into urban areas because of illness. Using Danish whole-population data, it could be shown that urban upbringing itself was associated with greatly increased rates of psychosis in later life (Pedersen 2001). By measuring the exposure during childhood, rather than adulthood as previous studies had done, a causal path could be more clearly established.

A key component of this kind of population registry data is that every citizen has a unique personal identification number, which is included in all their official records. This makes it possible to easily link individual health records over time and to link data across a wide range of different domains. For example, records for psychiatric in-patient stays can be linked to out-patient appointments, medication use, tax and employment records, migration and educational data (Pedersen 2001; Norredam 2011; Schofield 2017a). They can also be linked to blood samples from which it is possible to extract DNA for genetic research (Agerbo 2015).

Scandinavian countries are not alone in making population health records available for research. A recent report from the Organisation for Economic Co-operation and Development (OECD) also highlighted Korea, Singapore, Israel, New Zealand and the UK as scoring highly on the availability of population health data for research (OECD 2015). However, Scandinavian countries have the advantage that, because these data have been collected in electronic form since the 1960s, it is now possible

Peter Schofield is a Medical Research Council research fellow in the School of Population Health & Environmental Sciences, King's College London. His research uses a mixed-methods approach, including analysis of whole-population data, to investigate the role of social factors in the aetiology and management of mental disorders. **Jayati Das-Munshi** is an honorary consultant psychiatrist with South London and Maudsley NHS Foundation Trust. She also holds a clinician scientist fellowship from the Academy of Medical Sciences/Health Foundation and is based at the Institute of Psychiatry, Psychology & Neuroscience, King's College London. Her areas of interest include the social determinants of mental disorders, including ethnic minority/migrant health inequalities, the interplay of physical and mental health, and novel methodologies to address research questions.
Correspondence Peter Schofield, School of Population Health & Environmental Sciences, Faculty of Life Sciences and Medicine, King's College London, 3rd Floor, Addison House, Guy's Campus, London SE1 1UL. Email: peter.1.schofield@kcl.ac.uk

to access population cohort data over much of the life course (Rosen 2002).

We are moving towards developing similar resources that are applicable to mental health research in the UK. These include the Scottish health and ethnicity linkage study (SHELS) (Bhopal 2011) (see below), primary care data such as the Clinical Practice Research Datalink (CPRD), and linked psychiatric case register databases such as the Clinical Records Interactive Search (CRIS) (also discussed below). Attempting linkages across administrative and health records without an equivalent universal personal identification number can, however, be methodologically challenging.

What can big data in mental health really achieve?

Increasingly, new resources are being created with exciting possibilities in terms of their potential application to mental health research in the UK (McIntosh 2016). However, as we have highlighted, large-scale electronic data resources have existed for many decades in other settings. Some of the methodological advances made using these resources can inform how we progress with the resources that have become available in the UK.

Big data can facilitate novel study designs

Some would argue that the gold standard for evidence is the well-conducted randomised controlled trial (RCT). Yet, in many situations it is challenging or even impossible to conduct this type of study. For example, for rare outcomes such as suicide following an episode of self-harm, it can be difficult to ensure that sample sizes are adequate or that the follow-up period is long enough to detect the possible beneficial effects of an intervention.

In this context, large routine electronic datasets may help in assessing which types of intervention are beneficial, even if an RCT is not possible. For example, Erlangsen and colleagues (2015) used whole-population data from Denmark to assess the role of a psychosocial intervention in reducing subsequent completed suicide risk in a national sample of people who had self-harmed. The authors used propensity scores – an approach that can be used in observational data that involves matching on variables that predict outcomes, which leads to a replication of the balance that is normally achieved in well-conducted RCTs. By achieving this balance in a cohort of individuals who had self-harmed (with data on 5678 individuals who had received the psychosocial therapy following a self-harm episode and 17 034 individuals who had not) the investigators were able to establish that receiving a psychosocial intervention that focused on suicide prevention after an initial episode of self-harm reduced the risk of repeated self-harm episodes and death by any cause 1 year after the index event and was also associated with a reduced risk of repeated self-harm, death by suicide and deaths by any cause 5–20 years after the intervention.

Studies such as this are clearly a powerful example of how routine electronic data on large samples might be applied to contexts where standard RCTs may not be feasible. Erlangsen and colleagues do caution that, despite the sophisticated methodologies employed, selection bias, as well as a lack of more detailed information on what the 'psychosocial therapy' entailed, are potential limitations for their study (Erlangsen 2015), yet work like this highlights the possibilities of big data in making important contributions to the mental health evidence base.

Big data can enhance study recruitment

Recruitment to clinical trials can be difficult, with specific challenges relating to the recruitment of people with mental disorders (Howard 2009). This may partly be a function of clinicians acting as gatekeepers, to the extent that patients may not be offered the opportunity to participate in studies, even if they wish to do so (Callard 2014; Patel 2017). Details of one innovative system to enhance the recruitment of patients into research studies, which was developed in partnership with patients and carers and is based on an anonymised electronic health record system, is highlighted Box 1. Despite the challenges outlined in Box 1, the 'consent for contact' (C4C) system is an extremely innovative example of what may be possible using large electronic health record resources, with models developed in partnership with patients and carers.

Big data can enhance RCTs

Embedding well-designed RCTs within everyday clinical practice, where electronic health records have fully replaced paper-based systems for medical note-keeping (Gulliford 2014; McIntosh 2016), so that there can be 'randomisation at point of routine care' (van Staa 2012) is another innovative study design that is yet to be fully realised in psychiatry. As mental health trusts increasingly move towards fully electronic medical records, such a 'mixed design' which intermeshes the clinical trial with automated data collection (including data on outcomes such as adverse events, has obvious logistic and cost advantages. For example, potential participants for research trials may be followed up through the data routinely noted on the electronic health record, which may be of particular value for adverse events (van Staa 2012). This type of study

BJPsych Advances (2018), vol. 24, 237–244 doi: 10.1192/bja.2018.15

BOX 1 Consent for Contact (C4C)

The 'consent for contact' (C4C) system in South London and Maudsley (SLaM) NHS Foundation Trust is an innovative example of a system whereby the autonomy of patients wishing to take part in mental health research is enhanced through a robustly anonymised electronic health record system (Callard 2014). This was developed with considerable patient and carer involvement and is based on the SLaM Biomedical Research Centre's CRIS register, comprising the fully de-identified health records system for a large mental health trust containing over 250 000 patient records over a catchment area of 1.2 million people (Perera 2016). In the C4C system, care coordinators or others in the patient's team are able to ask patients whether they would be willing to join the C4C register, through a consenting procedure that clarifies to the patient that they are joining a register where they may be contacted in future to take part in research (rather than giving consent for a specific research project) in a range of areas, with the patient able to refuse at any point (Callard 2014). Once a patient consents to join the C4C register, this is flagged on their electronic health record. With the numbers who have consented now in the thousands (Oduola 2017) this is an invaluable resource, as investigators may otherwise struggle to recruit hard-to-reach or underserved populations.

design has already been employed in trials of antibiotic prescriptions and stroke prevention (Gulliford 2014) and it might also be suitable for studies of mental health interventions.

Big data can enhance health records through data linkage

Unlinked data from health records may be missing important information, which could hamper analyses. Frequently, important indicators of health outcomes and important sociodemographic variables such as ethnicity are poorly recorded or of variable quality (Bhopal 2011). Linked datasets (Box 2) allow the possibility of bringing in information from various sources to create large cohorts or datasets of individuals with less common conditions on a scale that is difficult to otherwise achieve in traditional epidemiological studies. This is partly because traditional epidemiological studies may be hampered by challenges of recruitment, loss to follow-up/attrition and falling participation rates (Knudsen 2010). The linkage of data to routine sources additionally helps to 'plug the gap' if important indicators, such as self-ascribed ethnicity, can be brought in via the linkage (Bhopal 2011). For example, a linkage of health data to census records in Scotland (SHELS) highlighted ethnic minority mental health inequalities specific to the devolved Scottish context (Bhopal 2011; Bansal 2014). Traditional studies using unlinked routine data would not have been able to achieve this, as ethnicity was not routinely recorded in Scottish health records at the time of the study. In England, the linkage of death certificate information to records from mental health trusts have informed our understanding of premature mortality in severe mental illness (Chang 2011; Das-Munshi 2017) as well as conditions such as chronic fatigue syndrome (using the CRIS register; Roberts 2016). The linkage in these examples enabled a sample size that allowed sufficiently powered analyses.

What big data *cannot* do

Although this kind of big data clearly has enormous advantages for the type of research that we do now and that will be possible in the future, it is very easy to lose sight of some of the inherent limitations that come with these resources.

BOX 2 Examples of linkages and clinical applications

Death certificate information linked to electronic health records

In a study of severe mental illness (SMI), the investigators linked electronic health records from a large case registry from a mental health trust in London to death certificate information. This study highlighted a substantially lower life expectancy in people with SMI, with the greatest reduction in men with schizophrenia (14.6 years lost) and women with schizoaffective disorders (17.5 years lost) (Chang 2011).

National pupil database linkage to mental health records

A recent linkage of mental health data with data from the national pupil database has allowed the possibility of bringing together clinical mental health data and teacher-assessed measurements of developmental and special educational needs from the schools' database (Downs 2017). The basis of this linkage in real-time electronic health records has the potential to inform service development and be used as a tool to monitor and evaluate service improvements.

Primary care linkage to mental health records

It is a concern that people with SMI experience premature mortality, with most deaths from preventable physical causes such as cardiovascular disease. The 2012 National Audit of Schizophrenia, covering England and Wales, revealed low levels of recording of physical health indicators such as body mass index (BMI) in people with SMI (Crawford 2014). In the UK, most people are registered with a GP/family doctor in primary care, which is where most physical healthcare is monitored and recorded. Therefore, in the UK, linkage of primary care records to secondary mental healthcare records can shed light on the quality of physical healthcare received by people with SMI. For example, in a study that used such a linkage, the investigators found that people who had coronary heart disease or heart failure comorbid with SMI were more likely to receive suboptimal treatments for these conditions than those without comorbid SMI (Woodhead 2016). This was especially the case in individuals prescribed depot antipsychotic medications, in those identified as having SMI of greater severity and in those with one or more recorded risk events.

Big data cannot replace statistical analysis

A common misconception when presented with data collected for the entire population is that we no longer need to be concerned about the statistical significance of our findings. Statistical theory presupposes that the data we are analysing can be treated as a random sample of the overall population of interest. Therefore, it is often assumed that if we know the health outcomes for the entire population then we no longer have to worry about burdensome statistical calculations. We could instead simply give the percentage of people with, say, a diagnosis of schizophrenia who were exposed to some risk factor and the percentage not exposed and assume that this covers everything. However, research is rarely about what has already occurred. Instead, we intend that research findings are relevant for future situations and allow us to develop some overarching theory. Even the most complete population data will only ever comprise a subset of all possible instances of the phenomenon of interest and therefore statistical methods are needed to account for this.

Big data cannot predict the future

This brings us to arguably the most common example of 'big data hubris': that the more data we collect the more likely we will be able to accurately predict future events. A good example of this, that is often cited, is the case of Google Flu Trends (Box 3). Although it appears that Google has abandoned this project, others believe that this is far from the end of the story. One wide-ranging review argues that similar algorithms could be successful, although they would require constant updating and improvement and should ideally be used alongside other epidemiological tools (Lazer 2014).

BOX 3 Google Flu Trends

Originally heralded as an exemplar of the use of big data, Google Flu Trends used patterns in large numbers of Google searches to predict localised flu outbreaks (Mayer-Schönberger 2013). By mining how combinations of search terms were related to subsequent outbreaks, the resulting algorithms were used to predict future epidemics. Initial success led to claims that Google Trends would ultimately replace costly epidemiological surveys. But this was short lived, as changes in the way that the Google searches were conducted and processed, and some of the underlying assumptions, led to overestimates of disease incidence that were no better than those based on historical data alone (Lazer 2014).

A telling legacy of this is the official website to which searches for 'Google Flu Trends' are currently directed (www.Google.org/flu-trends/about). Adopting a cheery tone, 'Thank you for stopping by', this documents how models were first developed in 2008 only to be discontinued in 2014, concluding that it is 'still early days for nowcasting'.

Big data cannot make up for the absence of theory

Along with the initial wave of enthusiasm about big data came the idea that 'text mining' of large datasets to find relevant patterns was methodologically valid in itself. Concerns about causality and the reasoning behind these algorithms were seen as irrelevant as long as the algorithms worked, as was the case with Google Flu Trends (Box 3) (Mayer-Schönberger 2013). This is the logic behind 'machine-learning' approaches. In machine learning a 'training' dataset is used to develop an algorithm from a large set of often arbitrary variables. This algorithm is then applied to another set of test data until an optimum predictive tool is arrived at. This kind of 'black box' approach is therefore essentially atheoretical – the authors do not need to know why the algorithm works, simply that it does work when applied to the test data. It is not hard to see how this could be attractive to mental health research, where many fundamental questions about aetiology remain unanswered. Instead of trying to determine the mechanism that might lead to, say, increased rates of schizophrenia among migrants, a simpler approach might be simply to arrive at a predictive tool by determining patterns in available data. In fact, in the case of Google Flu Trends, the algorithms themselves were never made public, so it was impossible to determine why they were ever successful in the first place, or why they subsequently underperformed. Although this approach has enormous advantages in many fields, for example in machine translation and text mining, it is highly problematic in epidemiological research, as Google Flu Trends demonstrated.

Big data cannot always be taken at face value

Unlike research data, administrative data seldom come with documentation explaining how the data were collected or how categories used in the coding were arrived at. For example, if we are interested in rates of mental disorder for different ethnic groups, with survey data we can determine whether ethnicity is self-reported or not, as well as the categories used in the original questionnaire. However, if we look at health records, in the UK, it is often impossible to say who provided the ethnic classification or how it was originally coded. This could be a problem if we tried to determine ethnic health differences between different areas, but were unable to distinguish between differences in coding methods and underlying health differences.

Often, the way that administrative data are presented suggests a completeness and objectivity that can be misleading if taken at face value. For example, just because a field is presented in general

BJPsych Advances (2018), vol. 24, 237–244 doi: 10.1192/bja.2018.15

practitioner (GP) data for a diagnosis of depression this does not mean that it can be useful if we wish to determine prevalence (Box 4). It is possible, however, in some instances to use a hybrid screening approach to make up for this. For example, for rare disorders such as psychosis more accurate diagnostic coding has been achieved using a combination of clinical expertise and machine-learning techniques to process detailed data from health records documenting symptoms (Patel 2015; Gorrell 2016). Often, however, we do not have detailed symptom data. It is therefore important to be aware that all data are created in a context, whether social, administrative, technical or clinical. If we fail to take this into account, we risk misinterpreting the data we have collected (Hennekens 1987; Prince 2003).

BOX 4 Depression coding in GP records

The way that data are coded can reflect administrative priorities that are at odds with research. For example, for some time the way that depression diagnosis has been coded in National Health Service primary care data has meant that it is underrecorded compared with what we know from national surveys (Rait 2009; Kendrick 2015). Much of this has been a result of changes (in 2004) to the way that GPs are incentivised. These changes mean that recording a diagnosis of depression can lead to triggers in the clinical record for further action to be taken that many GPs see as unnecessarily burdensome and not directly relevant to clinical care. Therefore, many GPs simply enter a different term in the record, such as 'low mood'. Although this did not affect clinical care, it led to an underestimate of the prevalence of depression in primary care, as the 'low mood' term is not captured by diagnostic systems. Therefore, without understanding what statisticians call the 'data generating mechanism' behind this kind of health records data, it would be easy to misinterpret what appears to be very low prevalence.

Big data alone cannot solve complex analysis problems

Large datasets of population health records can help solve one of the major challenges of psychiatric research, by providing adequately powered samples of the population of interest. However, the challenges of data analysis do not become easier simply because more data are collected. In fact, the larger the dataset the greater the potential complexity to be accounted for in the analysis. With a small well-designed trial or survey potential confounding variables, i.e. patterns in the data that could obscure our findings, can often be easily accounted for. However, population health records do not come with any such safeguards and are easily open to misinterpretation due to our failure to account for these patterns. For example, we could misinterpret spatial patterns by failing to account for differences in contextual risk factors such as urbanicity (see above), as well as differences in the reporting practices of mental health trusts in different parts of the country. Similarly, temporal patterns could also confound our results, such as changing ICD diagnostic categories. For example, with the change from ICD-9 to ICD-10 the latter showed a much higher sensitivity for dementia, which could easily be misinterpreted as an increase in prevalence if we were examining trends using health records data alone (Quan 2008).

To account for this often requires a quite different analysis approach to the statistical methods used with smaller, more theoretically determined, samples. Where a well-designed RCT could potentially be analysed using routine techniques such as a *t*-test, for whole-population data more complex multilevel modelling and Bayesian analysis are often necessary. Although this is becoming easier with the widespread adoption of more advanced statistical methods, these still remain beyond the expertise of many researchers.

Big data cannot make research more replicable

In recent years, increasing concern has been raised about the 'replicability crisis' in scientific research, and particularly psychological research. For example, in one recent poll of 1500 scientists 70% had failed to reproduce another scientist's experiment and around 50% had failed to reproduce one of their own experiments (Baker 2016). Often, examples are given of small-scale psychology experiments yielding interesting findings that consistently fail to be replicated. It could be argued that big data is one solution to this problem, as more data means results are more generalisable and therefore replicable. However, this is to misunderstand the nature of the problem. It is typically not the size of the dataset that is at issue, but the potential for spurious results in those situations where the researcher is faced with a multitude of different possible interpretations of the data. As datasets become larger and more complex, the number of potential subgroups to be analysed, analysis methods used and alternative categorisations to be adopted increases exponentially. For the unscrupulous researcher this could mean simply re-running the analysis by trying every possible combination of these until the results fit the required statistical significance (or '*P*-value') – a practice known as p-hacking (Gelman 2014). This has reached the point where the American Statistical Association (ASA) recently felt compelled to issue a formal statement about the correct use of *P*-values (Wasserstein 2016). For the ASA, the recent expansion in the use of large complex datasets for research, while expanding the possibilities for novel research, increases the risk of erroneous conclusions being made from the data. This may not even be deliberate; it is possible that, faced with many different analysis possibilities, the researcher may, whether consciously or not, be

MCQ answers
1 e 2 c 3 d 4 b 5 e

inclined towards the one more likely to give the desired result given the data they are presented with. It is very difficult to rule this out, although one solution is to make the analysis process more transparent (Box 5). For some types of research this can be achieved by reporting in advance the protocol for future studies, along with details of the analysis method. However, this is not necessarily applicable or helpful for many descriptive studies, where the ultimate focus may not be predetermined.

Big data cannot answer questions for which data have not already been collected

Big data is, by definition, data collected for purposes other than research and therefore does not always fit the research questions we wish to ask. For example, with diagnoses recorded over time for the purposes of clinical care, and not aetiological research, it is often very difficult to determine exactly how date of diagnosis relates to onset. Similarly, if we rely on big data alone it becomes very difficult to do research on disorders that have not already come to the attention of mental health services. In such situations, a reliance on routine health-systems data may under- or overestimate the actual prevalence of mental disorders. For such situations, cross-sectional surveys based in the community still have a major role to play. So, while traditional methods such as surveys, RCTs and qualitative studies allow us to determine what data are collected, with big data there is a danger that we neglect those research topics for which we do not already have available data (Schofield 2017b). This has particular relevance to mental health research, where social factors are often inextricably linked to the aetiology and progress of mental disorders (van Os 2010; Reininghaus 2014). A reliance on big data alone risks a circularity in the way research is conducted, as studies framed within a biomedical model, using data collected from medical records alone, remove the possibility that social factors might be included in the aetiology of mental disorder.

BOX 5 Ensuring transparency

Information from large-scale electronic health records has an important role to play in mental health research. However, as we have highlighted, there are major caveats to how the data are utilised when attempting to answer challenging questions related to mental health research. All research, including the best-designed studies, will have limitations. The reporting of research methods can be strengthened and made more transparent by adhering to principles advocated in guidelines such as STROBE (Strengthening the Reporting of Observation Studies in Epidemiology) (von Elm 2008) and CONSORT (Consolidated Standards of Reporting Trials) (Schulz 2010). Guidelines for the reporting of observational studies using routinely collected data have also been developed (Benchimol 2015). Adhering to guidelines such as these will ensure that studies conducted on electronic health records and other administrative data resources for mental health research are transparent and more likely to be replicable.

Conclusions

There are clearly considerable advantages to the use of 'big data' available in large health records' datasets for psychiatric research. However, these data sources come with inherent limitations, as we have outlined, and therefore should not replace methods for which there is already proven utility. Instead, we argue, they should play a complementary role alongside RCTs, representative surveys, cohort studies and qualitative studies, capitalising on the methodological advantages of each while offsetting their respective limitations. We have also outlined ways in which novel methodologies, such as quasi-experimental designs and embedded RCTs, as well as novel recruitment possibilities, may be intermeshed with big data to enhance traditional research methods. We are confident that many more such novel applications and methods will become apparent with time, as this field is rapidly changing.

References

Agerbo E, Sullivan PF, Vilhjálmsson BJ, et al (2015) Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: a Danish population-based study and meta-analysis. *JAMA Psychiatry*, **72**: 635–41.

Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature*, **533**: 452–4.

Bansal N, Bhopal R, Netto G, et al (2014) Disparate patterns of hospitalisation reflect unmet needs and persistent ethnic inequalities in mental health care: the Scottish health and ethnicity linkage study. *Ethnicity & Health*, **19**: 217–39.

Benchimol EI, Smeeth L, Guttmann A, et al (2015) The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Medicine*, **12**: e1001885.

Bhopal R, Fischbacher C, Povey C, et al (2011) Cohort profile: Scottish health and ethnicity linkage study of 4.65 million people exploring ethnic variations in disease in Scotland. *International Journal of Epidemiology*, **40**: 1168–75.

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

Chang C-K, Hayes RD, Perera G, et al (2011) Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London. *PLoS ONE*, **6**: e19590.

Crawford MJ, Jayakumar S, Lemmey SJ, et al (2014) Assessment and treatment of physical health problems among people with schizophrenia: national cross-sectional study. *British Journal of Psychiatry*, **205**: 473–7.

Das-Munshi J, Chang C-K, Dutta R, et al (2017) Ethnicity and excess mortality in severe mental illness: a cohort study. *Lancet Psychiatry*, **4**: 389–99.

Downs J, Gilbert R, Hayes RD, et al (2017) Linking health and education data to plan and evaluate services for children. *Archives of Disease in Childhood*, **102**: 599–602.

Erlangsen A, Lind BD, Stuart EA, et al (2015) Short-term and long-term effects of psychosocial therapy for people after deliberate self-harm: a register-based, nationwide multicentre study using propensity score matching. *Lancet Psychiatry*, **2**: 49–58.

BJPsych Advances (2018), vol. 24, 237–244 doi: 10.1192/bja.2018.15

Gelman A, Loken E (2014) The statistical crisis in science. *American Scientist*, **102**: 460.

Gorrell G, Oduola S, Roberts A, et al (2016) Identifying first episodes of psychosis in psychiatric patient records using machine learning BT. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* (ed Association for Computational Linguistics): 196–205. ACL.

Gulliford MC, van Staa TP, McDermott L, et al (2014) Cluster randomized trials utilizing primary care electronic health records: methodological issues in design, conduct, and analysis (eCRT Study). *Trials*, **15**: 220.

Hennekens C, Buring J, Mayrent S (eds) (1987) *Epidemiology in Medicine*. Lippincott Williams and Wilkins.

Howard L, de Salis I, Tomlin Z, et al (2009) Why is recruitment to trials difficult? An investigation into recruitment difficulties in an RCT of supported employment in patients with severe mental illness. *Contemporary Clinical Trials*, **30**: 40–6.

Kendrick T, Stuart B, Newell C, et al (2015) Changes in rates of recorded depression in English primary care 2003–2013: time trend analyses of effects of the economic recession, and the GP contract quality outcomes framework (QOF). *Journal of Affective Disorders*, **180**: 68–78.

Knudsen AK, Hotopf M, Skogen JC, et al (2010) The health status of non-participants in a population-based health study. *American Journal of Epidemiology*, **172**: 1306–14.

Lazer D, Kennedy R, King G, et al (2014) Big data. The parable of Google Flu: traps in big data analysis. *Science*, **343**: 1203–5.

Mayer-Schönberger V, Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

McIntosh AM, Stewart R, John A, et al (2016) Data science for mental health: a UK perspective on a global challenge. *Lancet Psychiatry*, **3**: 993–8.

Norredam M, Kastrup M, Helweg-Larsen K (2011) Register-based studies on migration, ethnicity, and health. *Scandinavian Journal of Public Health*, **39**: 201–5.

Oduola S, Wykes T, Robotham D, et al (2017) What is the impact of research champions on integrating research in mental health clinical practice? A quasiexperimental study in South London, UK. *BMJ Open*, **7**: e016107.

OECD (2015) *Health Data Governance: Privacy, Monitoring and Research*. OECD Publishing.

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

Patel R, Oduola S, Callard F, et al (2017) What proportion of patients with psychosis is willing to take part in research? A mental health electronic case register analysis. *BMJ Open*, **7**: e013113.

Pedersen CB, Mortensen PB (2001) Evidence of a dose-response relationship between urbanicity during upbringing and schizophrenia risk. *Archives of General Psychiatry*, **58**: 1039–46.

Perera G, Broadbent M, Callard F, et al (2016) Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open*, **6**: e008721.

Prince M, Stewart R, Ford T, et al (eds) (2003) *Practical Psychiatric Epidemiology*. OUP.

Quan H, Li B, Duncan Saunders L, et al (2008) Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Services Research*, **43**: 1424–41.

Rait G, Walters K, Griffin M, et al (2009) Recent trends in the incidence of recorded depression in primary care. *British Journal of Psychiatry*, **195**: 520–4.

Reininghaus U, Morgan C (2014) Integrated models in psychiatry: the state of the art. *Social Psychiatry and Psychiatric Epidemiology*, **49**: 1–2.

Roberts E, Wessely S, Chalder T, et al (2016) Mortality of people with chronic fatigue syndrome: a retrospective cohort study in England and Wales from the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Clinical Record Interactive Search (CRIS) Register. *Lancet*, **387**: 1638–43.

Rosen M (2002) National health data registers: a Nordic heritage to public health. *Scandinavian Journal of Public Health*, **30**: 81–5.

Schofield P, Das-Munshi J, Becares L, et al (2017a) Neighbourhood ethnic density and incidence of psychosis – First and second generation migrants compared. *European Psychiatry*, **41**: S249.

Schofield P (2017b) Big data in mental health research – do the *n*s justify the means? Using large data-sets of electronic health records for mental health research. *BJPsych Bulletin*, **41**: 129–32.

Schulz KF, Altman DG, Moher D, et al (2010) CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, **340**: c332.

van Os J, Kenis G, Rutten BP (2010) The environment and schizophrenia. *Nature*, **468**: 203–12.

van Staa T-P, Goldacre B, Gulliford M, et al (2012) Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ*, **344**: e55.

von Elm E, Altman DG, Egger M, et al (2008) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, **61**: 344–9.

Wasserstein RL, Lazar NA (2016) The ASA's statement on *p*-values: context, process, and purpose. *American Statistician*, **70**: 129–33.

Woodhead C, Ashworth M, Broadbent M, et al (2016) Cardiovascular disease treatment among patients with severe mental illness: a data linkage study between primary and secondary care. *British Journal of General Practice*, **66**: e374–81.

MCQs

Select the single best option for each question

1 A major advantage of big data in mental health research is:
a we no longer need to use statistics
b analysis is much simpler
c it is easy to get the results we want
d we no longer need other more expensive forms of research data
e we can answer many research questions previously beyond the scope of research.

2 When analysing big data, research questions are:
a no longer important
b easily matched with available data
c often outside the scope of available data
d the last thing we need to think about
e decided by the computer algorithm.

3 Big data cannot:
a be combined with other kinds of research data
b be used in experimental studies
c be used in study recruitment
d be interpreted without understanding the data generating mechanism
e be used for purposes other than that for which it was collected.

4 Big data can:
a allow us to predict what happens in the future
b improve our ability to make causal inferences
c replace the need to make causal inferences
d replace most other research resources
e do away with the need for epidemiologists and medical statisticians.

5 Big data is:
a something that has only existed in the past couple of decades
b not found in the UK
c confined to social media, e.g. analysing Facebook 'likes' and Twitter feeds
d a passing fad that serious researchers should ignore
e a major opportunity for enhancing mental health research.

BJPsych Advances (2018), vol. 24, 237–244 doi: 10.1192/bja.2018.15