

Data science involves working with large data sets, containing information from many hundreds and thousands of people – sometimes millions. These large data sets may be drawn from cohort studies, where participants have chosen to enrol and given explicit permission for their data to be used. They may also deploy routine data – information that is collected by health services for example, such as drug prescriptions or hospital admissions. We can use these routine data to understand long-term trends in mental health, and – crucially – to capture something of the experience of those people who don't enrol in cohort studies or struggle to stay in touch.

Mental health data science involves those with experience of mental illness, or who may be at increased risk of developing mental ill-health in the future. It is different from data science for physical conditions (such as cancer) because of:

- Stigma, which can still be associated with mental ill-health
- The rich qualitative data available, such as routinely collected therapist case notes
- Lack of trust/dissatisfaction with health services, that some people with mental ill-health express

Together, these factors may make people with mental illness feel less confident about data science, and more likely to object to researchers accessing routine data without informed consent.

What can we do about this?

We believe it is essential to engage with stakeholders and use their perspectives to inform how mental health data science is done.

To that end we recruited an expert task force of about 30 people who all had lived experience of mental illness and professional knowledge about data science. Over three rounds, in early 2020, they helped us generate, refine and then finalise this checklist which aims to define **best practice in mental health data science** from a well-informed, lived experience perspective. The wording overleaf represents the distillation of their combined points of view.

The best practice checklist comes in two versions. The first version describes what researchers can do **today** to ensure that their work is respectful and trustworthy. The second version describes what we, as a mental health data science community, should deliver **in the future** for excellence in mental health data science.

Find out more

You can find out more about the checklist and its creators, and read the academic articles that underpin this output, at www.mhdss.ac.uk.

To cite this checklist please use:

Kirkham, E.J., Iveson M., Beange, I., Crompton, C.J. McIntosh, A. & Fletcher-Watson, S. (2020) A stakeholder-derived, best practice checklist for mental health data science in the UK. University of Edinburgh.

Thank you

We would like to thank the Delphi experts who co-produced these guidelines, the MRC Pathfinder Stakeholder Advisory Group at the University of Edinburgh, our MRC Pathfinder colleagues both within the University of Edinburgh and across the UK, and all those who helped us to recruit our panel of experts.

Best Practice for Mental Health Data Science

This checklist was co-created by community experts with a research team. Its purpose is to guide the practice of data science in mental health research.

How to use the checklist

The checklist is designed to complement other guidelines for data science practice, such as the UK Government's Data Ethics Framework. It gives researchers information about what is acceptable to, and important for, people with lived experience of mental illness.

Researchers conducting new studies should consider each item on the **today** checklist and put measures in place to meet these best practice criteria. Together as a research community, we should also aim to develop and disseminate new protocols to meet the criteria for best practice **in the future**.

Our website at www.mhdss.ac.uk includes case studies which illustrate how different mental health data science teams around the UK are delivering excellence in their research practice.

Best practice for mental health data science **today** means...

Best practice for mental health data science **in the future** means...

Best practice for mental health data science **today** means...

Best practice for mental health data science **in the future** means...

Security

...data should be accessible to a range of people who conduct research, including authorised academics and clinicians.



...providing appropriate training and supervision for data users, and carrying out criminal record checks where relevant.

...ensuring that data are accessed in safe settings using clear and efficient procedures.



...providing digital controls to allow remote access from private settings, using procedures that are robust and easy to follow.

...creating data management plans and ensuring that these are adhered to at all times.



...incorporating inspection processes to ensure ongoing compliance with good data practice, and responding proportionately to inappropriate data use with measures such as training or temporary or long-term suspension of access.

...planning in advance to prevent data breaches, using a recording process for data breaches, and reporting near misses.



...developing robust systems to prevent data leaks and record data breaches and near misses.

...using de-identified data, except where identifiable information (including information about protected characteristics) is essential to beneficial outcomes. In all cases the health and benefit of people with lived experience should be prioritised.



...building plans for de-identified data collected by researchers, scientists and clinical services to be made available for analysis on an open-access basis.

Anonymity

...researchers, scientists and clinical services making de-identified data and findings (including null results) open-access where possible. It also involves awareness of the risk that qualitative data (such as free text) could contain identifiable information.



...developing effective measures, including secure linking systems, to protect against inappropriate identification and misuse.

...incorporating statistical disclosure control by following rules designed to prevent identification of individuals.



...incorporating statistical disclosure control based on principles, such as the principle that no individual may be identified, with training and external oversight.

Security

Anonymity

Transparency

Community

...ensuring that researchers have a process in place for responding to withdrawal requests and that they provide transparency on whether, how and when participants can withdraw their data.



...appointing a qualified, independent arbiter to arbitrate on complex questions relating to consent and data withdrawal.

...allowing other researchers to check analyses wherever possible (in addition to peer review).



...developing methods for de-identification, including innovative ways to mask identifiable information.

...monitoring data quality and taking account of the origin and quality of data when drawing conclusions.



...incorporating oversight of data repositories in order to monitor data quality and respond to public enquiries.

...facilitating research linking mental health data with other sources of public data, such as education or welfare data, in order to provide new information of public benefit about mental health.



...providing access to synthetic data where real data cannot be shared, in order to allow other researchers to check analyses and conclusions.

...ensuring that participants have as much control over consent as possible.



...exploring alternative models of consent, which may involve moving away from individualised models of consent.

...incorporating the views of people with lived experience throughout the course of each project, and providing nuanced and high-quality public communication of findings.



...incorporating the views of people with lived experience throughout the course of each project. At the same time the principles of open access should be followed by publicly pre-registering studies and providing accessible online information of each overarching request to use data, each output, and any null results.

...ensuring that data users understand the underlying data collection tools as well as the socio-cultural context in which studies are designed and findings are disseminated.



...active commitment to reducing the stigma associated with mental illness and its research, and to increasing public understanding of science.

Transparency

Community