Research Inclusion: Position statement on the use of external datasets

Purpose of this Statement

This statement sets out the NIHR Biomedical Research Centre (BRC): Oxford Health's position on the ethical and scientific importance of using data, including biological samples, that meaningfully reflect population diversity. It applies specifically to the use of externally sourced datasets and biobanks where representativeness is shaped by historical, structural, and practical factors beyond the control of our research teams. The purpose of this statement is to support transparent context-specific consideration of limitations and responsible interpretation of findings. In conjunction with the NIHR-INCLUDE (Innovations in Clinical Trial Design and Delivery for the Under-served) Framework and other relevant policies, it is intended to guide researchers, to ensure that our research contributes to equitable and inclusive health outcomes and delivers findings that are representative of the communities that we serve.

Importance of Representation and Inclusion in Data

NIHR BRC: Oxford Health is committed to producing research that benefits all communities. This requires using data and biological samples that reflect the diversity of the populations affected by health conditions we study. Diverse and inclusive datasets help ensure:

- Research findings are meaningful and applicable to the full spectrum of individuals who experience the health challenges that are the focus of the research.
- Scientific insights and outputs (for example medical interventions) do not cause harm, inadvertently perpetuate health inequalities or create new ones by over-representing specific groups and under-representing others.
- Research does not disproportionately reflect those who are already more likely to participate (e.g., white, more affluent, more educated groups), which is a recognised pattern in many existing large UK biobanks and datasets.

Current Limitations in the Data We Use

Much of our research relies on externally sourced datasets. While these resources are invaluable, they may lack demographic diversity. This issue is widely recognised, with increasing calls for structural changes in how externally available datasets including biobanks collect, curate, and govern data.

Limitations in external datasets vary by research context, health condition, and method. Common examples described in the literature include (illustrative, not exhaustive):

- **Ethnicity and ancestry**: Despite two decades of research, genome-wide association studies remain overwhelmingly based on participants of European ancestry, restricting the relevance and transferability of findings for ethnically diverse populations (Fitipaldi & Franks, 2023)[i].
- **Socioeconomic, disability, and health:** Analyses of UK Biobank data demonstrate that individuals with poorer health, chronic illness, and mental health difficulties are less likely to participate, meaning that populations with functional limitations and disability-related characteristics are under-represented even in the absence of explicit exclusion criteria (Schoeler et al., 2023)[ii].

Using this Position Statement

In selecting and using external datasets researchers should:

- Consider the most appropriate dataset for their work in the context of optimising inclusivity and representativeness.
- When writing up for publication, include a concise description of relevant limitations in sample representation; the potential impact of the above; any steps taken to mitigate, contextualise, or justify these limitations within the study design, analysis, or reporting.
- Consider, where there are sufficient numbers, preliminary analyses of differences across groups in the phenomenon under study

To accompany the above, researchers may use the below excerpt and should reference this full statement.

> NIHR Biomedical Research Centre (BRC): Oxford Health recognises that some externally sourced datasets do not fully represent the diversity of populations affected by the conditions we study. In accordance with good study practice, the NIHR BRC: Oxford Health supports researchers sourcing alternative external data, and other steps taken by researchers, to optimise inclusivity and representativeness in their analysis, and to describe any limitations in the writing up of the work.

Our aspirations and future commitments

This is an evolving area of work for NIHR BRC: Oxford Health. This position statement establishes direction and expectations and will inform the future development of guidance, tools, and resources. It does not constitute an action plan at this stage.

Additionally, NIHR BRC: Oxford Health will continue and seek to:

- **Strengthen Partnerships:** Work collaboratively with community organisations and minoritised health groups; for example, ethnic minoritised communities and people with intellectual and/or sensory disabilities. We will increase engagement with NHS partners and seek out national initiatives to improve representation in future collection of datasets, especially where these are designed for open use and/or available to external partners.
- **Explore Alternative Data and Sample Providers:** Identify and engage with biobanks or repositories that prioritise equitable recruitment, including among under-represented populations, even when this may be a smaller sample or incur an additional appropriate cost. We recognise that all not biobanks are equal and have differing access and ethics requirements.
- **Influence Sector-Wide Change:** Advocate within research networks, funding bodies and consortia for improved standards for demographic transparency, data provenance, and inclusion targets.
- **Develop Internal Guidance:** Create BRC-wide expectations for how researchers should acknowledge limitations in data representativeness and interpret findings responsibly.
- **Support Community-Involved Research Models:** Encourage approaches such as participatory design, community consultation, and co-creation, reflecting lessons from biobank governance about the importance of trust, engagement and culturally-sensitive practices.

Existing exemplars of good practice within NIHR BRC: Oxford Health themes:

Reviewing Measures and Tasks to Improve Data Quality: Depression Therapeutics Theme

The Depression Therapeutics theme provides an example of addressing data limitations by critically examining the tools used to generate commonly relied-upon research datasets. Through the *Your Voice Matters* project, supported by Public and Community Engagement with Research (PCER) funding (University of Oxford), the theme is working with people with lived experience of depression, researchers, and community partners to evaluate standard depression questionnaires. By exploring what these measures capture well, where they fall short, and how they are experienced by participants, the project recognises that limitations in data quality and interpretability can arise from the instruments themselves, not only from who is recruited. The work combines focus groups, including with the Oxford Health BRC Diversity in Research Group, and a large nationally representative survey, to produce clear guidance for researchers on responsible questionnaire selection, use, and communication. In

parallel, the theme has reviewed cognitive tasks to better reflect ethnic diversity in facial stimuli and age diversity in word stimuli, illustrating how adapting research measures can improve inclusivity and the relevance of resulting datasets.

Inclusive Scanning Project: Brain Technologies Theme

The Inclusive Research (Inclusive Scanning) project at the Oxford Centre for Integrative Neuroimaging demonstrates how limitations in research datasets can arise from standard research infrastructure and participant experiences, rather than from lack of willingness to participate. Established in 2022, the project critically examines how imaging methods, communication, consent processes, and research environments may exclude certain groups, contributing to datasets that are less diverse than the populations they aim to represent. By working collaboratively with researchers, radiographers, engagement specialists, and public contributors, the project has developed practical adaptations such as accessible “what to expect” resources and modified scanning processes. Importantly, the project has begun monitoring participant demographics to assess whether these changes are improving access and representation, illustrating how reflective infrastructure design and evaluation can help generate more inclusive and interpretable datasets.

Practical Checklist for Embedding Inclusivity in Behavioural Labs: Pain Theme

The Pain theme provides a complementary example of addressing data limitations through methodological reflection and co-production. Recognising that behavioural research datasets are often derived from narrow, unrepresentative samples, this work developed a practical checklist to help researchers identify and reduce barriers to participation in lab-based studies. Co-produced over three years with people living with chronic pain, neurodivergence, and caring responsibilities, the checklist prompts researchers to reflect on inclusion and exclusion at every stage of the research process, from study design and eligibility criteria, to data collection and dissemination. By encouraging adaptations for fatigue, mobility, sensory needs, scheduling, and communication, the checklist demonstrates how inclusive study design can directly improve data quality, relevance, and generalisability, particularly for groups most affected by health inequalities.

Addressing Representational Bias: Psychological Treatments Theme

The Psychological Treatments theme provides an example of addressing limitations in the representativeness of intervention materials in psychological treatments. Through a Programme Development Grant (NIHR) supporting the *Feeling Safe Programme* - a leading cognitive-behavioural intervention for persecutory delusions - the team undertook work to improve the treatment content to reflect further the diversity of people experiencing psychosis. The programme was revised to include new patient accounts from people with Black African and Black Caribbean backgrounds, expanded voice-over options, and another inclusiveness review conducted with lived-experience advisors.

This review involved 19 PPI consultants recruited across Black African, Black Caribbean, and Bengali communities, with representation across age and gender. The project demonstrates how co-production and content adaptation can improve inclusivity, acceptability, and relevance, supporting even more equitable application of evidence-based psychological treatments across diverse populations.

Integrating Diverse Data Sources for Personalised and Representative Evidence: Data Science Theme

The *Prescribing the Right Agent for Depression in Adults* study (PRADA) provides an example of addressing limitations in traditional research datasets by integrating multiple data sources to support more personalised and representative treatment decisions. The study recognises that standard approaches to antidepressant prescribing are often based on population averages that do not reflect individual variation. PRADA combines clinical, socio-demographic, and patient-reported data with genetic and digital measures to inform treatment selection, drawing on data collected across different healthcare settings and international populations. By linking and enriching datasets in this way the study demonstrates how research can move beyond narrow or homogeneous samples, supporting more inclusive, context-sensitive, and clinically relevant evidence for mental health care.

Improving Inclusion and Representativeness in Patient and Public Involvement, Clinical Trials and Complex Datasets: Preventing Multiple Morbidities Theme

The Preventing Multiple Morbidities theme contributes an overarching example of addressing limitations in patient and public involvement, clinical trial recruitment, and research datasets through *a critical examination* of how standard research designs and methodology systematically exclude people with complex intersecting health and social needs. Drawing on learning from NIHR and Research Council funded studies, this work shows that people living with multiple long-term conditions, severe mental illness, and social adversity are frequently under-represented because of rigid eligibility criteria, narrow survey categorisation, and inflexible research infrastructures. The paper highlights how participatory, co-designed, and adaptive research approaches can improve who is included in research and what data is collected, particularly for populations experiencing multimorbidity and health inequalities. By emphasising transparency about inclusion aims, flexibility in data collection, and critical reflection on commonly used categories and measures, this work demonstrates how research programmes can move beyond single-condition models to generate datasets that are more inclusive, interpretable, and relevant to real-world complexity.

Informed Selection of Data Sources to Improve Generalisability: Better Sleep Theme

The Better Sleep theme demonstrates good practice by actively reflecting on the strengths and limitations of the large population datasets it uses, recognising that no

single cohort is fully representative. By systematically considering diversity, ancestry, and inclusion across multiple biobanks and cohorts, the theme is able to combine and compare datasets in a way that improves generalisability and transferability of findings. This conscious, critical approach to dataset selection supports more inclusive and robust sleep research, despite the inherent limitations of individual resources.

Review and governance

This statement should be reviewed every **two years** or sooner if substantial changes occur in national policy, funding requirements, or the research environment.

This position statement has been written in consultation with:

- Theme EDI Ambassadors
- Theme Leads
- NIHR BRC: Oxford Health Patient and Public Involvement and Engagement Operations Group

This position statement has been reviewed and approved by:

- NIHR BRC: Oxford Health, Steering Committee

NIHR BRC: Oxford Health, Director, Professor Rachel Upthegrove

NIHR BRC: Oxford Health, Academic Equality, Diversity & Inclusion Lead, Dr Angharad de Cates

[i] Fitipaldi H, Franks PW. Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005-2022. Hum Mol Genet. 2023 Jan 13;32(3):520-532. doi: 10.1093/hmg/ddac245. PMID: 36190496; PMCID: PMC9851743.

[ii] Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. Nat Hum Behav. 2023 Jul;7(7):1216-1227. doi: 10.1038/s41562-023-01579-9. Epub 2023 Apr 27. PMID: 37106081; PMCID: PMC10365993.