Accessing large-scale health data for research: opportunities, barriers and risks

Yorkshire Cancer Research response, April 2026

Please briefly explain how your research expertise is relevant to the POSTnote topic.

Yorkshire Cancer Research has over 100 years' experience in funding and delivering research and services across Yorkshire, with £75.3 million currently committed to 59 pioneering programmes involving more than 750 researchers and cancer experts. This includes large scale clinical trials and community-based interventions that rely heavily on timely, accurate and granular health data.

The charity has developed a strong understanding of requirements involved in accessing, linking and analysing health datasets. Yorkshire Cancer Research is well placed to play a convening role, bringing together valuable insights from researchers across its diverse portfolio alongside evidence from its own data needs. This enables the charity to collate case studies and real world examples that illustrate the challenges researchers face within current data access systems.

What are the key issues relevant to the POSTnote that you would like to make us aware of?

Opportunities for large-scale health dataset research initiatives

The Health Data Research Service (HDRS) represents a significant opportunity to enhance the quality, reach and impact of health research across the UK. Allowing charities that fund health research and services to access HDRS datasets would further amplify this impact by enabling a broader range of evidence driven research and innovation.

At present, charities such as Yorkshire Cancer Research, are generally not considered for access to key data platforms such as CancerStats, despite having the facilities and security to access the highly valuable data they contain. These data could support the delivery of programmes such as Active Together, the Charity's pioneering prehabilitation and rehabilitation service for people with cancer, where access to timely, granular data on cancer related measures would enable more effective targeting, planning and delivery.

Improving access to data through the HDRS would also support a stronger understanding of regional need, allowing the Charity to identify priority populations, geographic inequalities and service gaps with greater precision. In turn, this would enable more strategic investment of research funding, more efficient service planning and ultimately improved outcomes for people affected by cancer in Yorkshire.

Case study: Yorkshire Lung Screening Trial (YLST)

The Yorkshire Lung Screening Trial (YSLT), funded by Yorkshire Cancer Research, ran from 2017 to 2025 and tested the impact of targeted lung screening in community settings, concentrating on deprived areas of Leeds. Evidence from this trial contributed to the UK National Screening Committee's recommendation for a National Lung Screening Programme.

YSLT used data from the National Disease Registration Service (NDRS) to collect long term follow-up outcomes and to apply data linkage for smoking cessation. Using NHS England Cancer Registration data, the study can compare stage distribution and numbers of late-stage cancers (stage III/IV) between intervention and control groups. This could show the true benefits or possible harms of introducing lung cancer screening nationally. Access to NDRS data for research is managed through NHS England's Data Access Request Service (DARS). The significant challenges encountered by the YLST team throughout this process provide important learnings for the development of largescale health data initiatives, including the forthcoming Health Data Research Service (HDRS).

Although the study's data use had already been approved by both the Health Research Authority Research Ethics Committee (REC) and Confidentiality Advisory Group (CAG), DARS required a separate approval process with its own advisory group. This duplicated scrutiny created avoidable delays and required large volumes of documentation, including certification from REC and CAG. It was not clearly communicated what documentation was needed for this process, resulting in significant back and forth between the two teams with DARS often requesting documentation they already had and requesting more documentation than was likely necessary for the process. Future platforms should have a more streamlined process with clarity from the outset on what documentation is needed with a central portal to store all required documents so that it is clear what has already been shared.

Further delays occurred when DARS applied updated data use standards during the trial. These changes resulted in requests to amend data uses that had already been reviewed and approved by REC and CAG. Longer studies like YLST are particularly vulnerable when regulatory requirements shift mid-project. There is a need for clear polices within future platforms to ensure that once data use is ethically and legally approved, it remains valid for the duration of the study unless significant new risks emerge.

The YLST team also experienced difficulties in determining what data was available within NDRS and how it was structured. Data was offered in packages or as individual variables, without guidance on what packages were available or which were suitable for their research needs. Meetings with analysts with in-depth knowledge of the data required waits of up to 16 weeks, and multiple meetings were often required to understand the impact of different data 'packages' or variable choices when building the data specification and then to help interpret the data. Although cleaned datasets were available, they were frequently up to 2.5 years old and lacked transparency about the cleaning processes. The lack of transparency and clear documentation led to delayed and incorrect submissions, resulting in repeated resubmissions. Comprehensive documentation and timely specialist support would allow researchers to submit accurate specifications at the outset.

More broadly, the data access process appeared not to have been designed with researchers' needs in mind. Platforms must be designed by, or in close collaboration with, experienced researchers to ensure that the process to access data is accessible, timely and fit for purpose. Researchers often apply data-minimisation principles, discounting data packages that include unnecessary variables. A lack of clarity on available datasets, timelines and suitability for specific purposes made it hard for the YLST team to adhere to these principles or plan work efficiently. DARS appeared to apply processes without proportionate consideration of the level of risk involved. These delays, compounded by poor communication, hindered the progress of the trial. A system designed to support research should prioritise timeliness to avoid unnecessarily stalling research. Future systems should embed proportionality as required by the

UK Policy Framework for Health and Social Care Research, with published reporting timelines and service standards to ensure accountability.

YLST experienced significant communication problems, including a lack of clear timelines for application processing, no updates on progress and uncertainty about points of contact. A transparent digital portal would significantly improve efficiency, allowing researchers to see required documentation and what has been submitted, real-time progress of applications and the most recent version of project specifications and data requests. Integrating a quality improvement feedback process would be valuable, enabling those who use the service to contribute to the development.