Uniting data for health equity

An ongoing question in public health is how best to identify problems contributing to inequity. For example, some populations living in a lower socioeconomic context are more likely to develop chronic illnesses such as type II diabetes. Similarly, a person's ethnic background can predispose them to various ailments, including diseases like cancer.

In theory, analysing health data can help researchers and clinicians to identify these risks at the group level. If we know, for example, that elderly people living in a particular part of the country are at a greater risk of respiratory illnesses, we can target interventions to those areas. We can personalize preventative practices to local populations — and in doing so, reduce the risk of many negative health outcomes, both physical and psychological, as well as the societal burden of diseases.

In practice, however, the absence of complete, representative datasets makes such projects difficult — and, in some populations, impossible. Datasets may be incomplete because data are not collected, or not provided by individuals or particular groups. This is especially problematic if these absences are more common in some groups than others. Our research suggests several reasons for these omissions, including technical problems such as with linking existing datasets that can help overcome absences, and social problems like persistently low trust in health and scientific institutions, especially among people with lower levels of education and historically oppressed groups.

These problems, which we detailed in recent and upcoming academic papers (Moorthie et al., 2022 and Moorthie et al., 2024), constitute a serious roadblock in our collective efforts to improve health equity through the effective use of data, and data-driven models. If we cannot see the specific health–related issues faced by particular populations, we cannot introduce preventive measures against these issues — putting the broader notion of data-powered public health in jeopardy.

There is room for optimism, though. Over the last year, we have consulted with colleagues working in healthcare in the East of England to develop case studies about how to address

issues related to data collection, maintenance, and linkage for the betterment of population health.

The first of these relates to ethnicity. We know that this characteristic is a critical element of understanding population-specific health-related risks. Ethnicity can indicate a likelihood of developing particular genetic diseases, such as Tay-Sachs disease in people of Ashkenazi Jewish descent, or it can help clinicians target measures, such as mental health interventions, towards historically mistreated populations. While 'ethnicity' has no universally accepted definition in the clinical or public health sciences, its use as a category in data collection and analysis is essential for population sub-groups facing shared challenges who may benefit as a group from targeted interventions.

Recognising this, the Performance and Analytics team at the Cambridgeshire and Peterborough NHS Foundation Trust (CPFT), who specialise in using data to identify risks to specific subgroups, learned that different clinical departments within their Trust varied in the quality of reporting and collecting ethnicity data. To improve reporting, they developed a dashboard so that teams could view their ethnicity data collection rates. They also ran workshops to train staff about the importance of ethnicity data collection.

Results were encouraging: while in 2022, 18% of patients seen had an ethnicity listed as 'not stated' in the Trust's databases, this figure decreased to less than 5% by the end of 2023.

Another initiative led by staff at the Suffolk and North East Essex Integrated Care System aimed to link datasets across the local region to help improve ethnicity data completions at the integrated level. Relying on a private technology firm, Optum, the project leads linked GP, Trust, and other local datasets, increasing ethnicity data completeness from 70% locally to nearly 94% across the area.

Both cases highlight the importance of senior-level staff buy-in, effective training, and improved dataset linkage and compatibility for improving the completion rates of a critical information (e.g. ethnicity) for combatting health inequities.

Our research highlights the need for two more important steps, essential for improving health equity across the country and beyond. First, we recommend establishing 'data coordinator'

positions working across trusts, GP services, and universities to drive improved data linkage and to design and deliver workshops training staff on the importance of accurate data collection. Establishing such roles would improve the completeness of recorded data, strengthening the robustness of analysis used to inform public health interventions. Moreover, it will save significant resources for different organisation as currently senior staff often have to dedicate significant amount of their time to dealing with administrative tasks associated with linking datasets.

For example, a 2025 New Zealand (Shabaz et al., 2025) study found that linking administrative datasets for a cohort of 859 participants required around 26 staff hours for coordination, with data access taking between 96 and 854 days and manual review needed for 6%–78% of records depending on dataset quality (Wiggins and Stokes, 2017). While no equivalent UK figures exist, these results serve as a useful benchmark for estimating the time and resources required for similar projects.

Second, we highlight the critical element of engaging with patients around their concerns about data collection and storage, with the hope that through effective policy and ethical practices, we can rebuild trust with the public — and especially those from groups with historic reasons for mistrust (Goodman and Milne, 2024; Goodman et al., 2025), for example the Roma.

We are currently working on a project exploring patients' preferences regarding how they place trust in people and institutions with their data, relying on a mixed methods approach using focus groups and a discrete choice experiment, which is a type of survey used to understand people's preferences on a particular topic. Through both projects, we aim to develop the groundwork necessary for building trust in healthcare institutions, with the hope that by doing so, we'll be able to help address the gaps in data that prevent us from finding the groups that need our help the most.

References

Moorthie, S., Hayat, S., Zhang, Y., Parkin, K., Philips, V., Bale, A., Duschinsky, R., Ford, T., & Moore, A. (2022). Rapid systematic review to identify key barriers to access, linkage, and use of local authority administrative data for population health research,

practice, and policy in the United Kingdom. *BMC Public Health*, 22, Article 1263. https://doi.org/10.1186/s12889-022-13187-9

Moorthie, S., Oguzman, E., Evans, S., Brayne, C., & LaFortune, L. (2024). Qualitative study of UK health and care professionals to determine resources and processes that can support actions to improve quality of data used to address and monitor health inequalities. *BMJ Open*, 14(9), e084352. https://doi.org/10.1136/bmjopen-2024-084352

Shahbaz, M., Harding, J. E., Milne, B., Walters, A., Underwood, L., von Randow, M., Jacob, L., & Gamble, G. D. (2025). Time and cost of linking administrative datasets for outcomes assessment in a follow-up study of participants from two randomised trials. *BMC Medical Research Methodology*, 25, Article 21. https://doi.org/10.1186/s12874-025-02458-9

Wiggins, N., & Stokes, B. (2017). Building a high quality data linkage spine using a targeted approach to clerical review. *International Journal of Population Data Science*, 1(1), Article 183. https://doi.org/10.23889/ijpds.v1i1.203

Goodman J.R., & Milne, R. (2024) Signalling and rich trustworthiness in data-driven healthcare: an interdisciplinary approach. *Data & Policy*. 6:e62. doi:10.1017/dap.2024.74

Goodman, J. R., Costa, A., & Milne, R. (2024). Trust in data sharing reflects the contextuality of the trustor–trustee relationship. *medRxiv*. https://doi.org/10.1101/2025.07.05.25330938