Ethnic bias in data linkage

In *The Lancet Digital Health*, Hannah Knight and colleagues[1] highlight stages in the data science pipeline that are affected by and lead to racism. Data linkage is a further stage in which ethnic bias can be encoded into datasets. Ethnic bias occurs when linkage error (false or missed matches) is more likely to occur for particular ethnic groups. The problem of ethnic bias in health data linkage is well described in the literature[2] and is concerning because health data are widely used for monitoring, service planning, research, evaluation, and policy. Systematic biases in data linkage misestimate health needs for ethnic minorities and further entrench existing disadvantages.

Accurate data linkage relies on accurately recorded identifying information and well designed linkage algorithms. However, ethnic minorities are more likely to have missing or incorrect information in their health records,[3] which might reflect structural biases in health systems (eg, ethnic minorities are more likely to be treated at health facilities with poorer overall data quality).[2] Data capture systems are also typically designed around Western name standards (ie, a first, middle, and last name) and do not account for cultural differences in name structures (eg, Hispanic groups can have multiple first or middle names, and often two surnames, and Asian names can follow different ordering norms). Linkage methods that require exact agreement on names can therefore contribute to ethnic bias. Requiring consent for linkage can also exacerbate bias, since ethnic minorities have higher rates of non-consent for linkage,[2] perhaps reflecting lower levels of trust in health systems and how their data are used.

Data providers and users should routinely explore ethnic bias by assessing data quality and linkage error.[4] Greater transparency of linkage processes, including routine reporting by disaggregated ethnic subgroups, would allow ethnic biases to be accounted for by statistical methods, and considered when assessing the validity of analyses and interpreting results. Data providers need to continually improve data quality and linkage methods (eg, through training of patient-facing staff in recording data for ethnic minorities, more inclusive data capture systems, and more flexible linkage algorithms). For example, we recently showed that, when linking administrative health and education records, relaxing requirements for exact matching on name improved linkage rates for ethnic minorities, although they remained disproportionately low.[5] Crucially, echoing Knight and colleagues,[1] we must all strive for greater diversity in the data linkage community, and more meaningful engagement with ethnic minorities to increase understanding of data linkage and address their concerns.

We declare no competing interests.

*Louise Mc Grath-Lone, Nicolás Libuy, David Etoori, Ruth Blackburn, Ruth Gilbert, Katie Harron
l.mcgrath-lone@ucl.ac.uk

Institute of Health Informatics (LMG-L, NL, DE,RB), Centre for Longitudinal Studies, Institute of Education (NL), and Great Ormond Street Institute of Child Health (RG, KH), University College London, London NW1 2DA, UK

1 Knight HE, Deeny SR, Dreyer K, et al. Challenging racism in the use of health data. *Lancet Digit Health* 2021; **3:** e144–46.

2 Bohensky MA, Jolley D, Sundararajan V, et al. Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010; **10:** 346.

3 Hagger-Johnson G, Harron K, Fleming T, et al. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open* 2015; **5:** e008118.

4 Gilbert R, Lafferty R, Hagger-Johnson G, et al. GUILD: guidance for information about linking data sets. *J Public Health (Oxf)* 2018; **40:** 191–98.

5 Mc Grath-Lone L, Blackburn R, Gilbert R. The Education and Child Health Insights from Linked Data (ECHILD) database: an introductory guide for researchers. London: University College London, 2021.