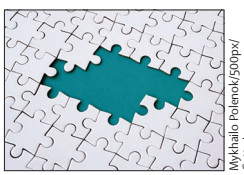
Lifting the veil on health datasets


Mykhailo Polenok/500px/Getty Images

As of Sept 22, 2024, SARS-CoV-2 has caused over 7 million deaths and new variants continue to emerge. Earlier this year, the WHO Coronavirus Network (CoViNet) was launched to enable early and accurate detection of coronaviruses and variant tracking. Public health monitoring systems like this are dependent on the availability and accessibility of global health data, but these data are often not representative of different demographic groups.

In this issue of *The Lancet Digital Health*, Joseph E Alderman and colleagues reviewed the composition and reporting of 192 publicly available health datasets that were used to train or test artificial intelligence (AI) algorithms during the COVID-19 pandemic. They identified substantial gaps in the metadata of these datasets, with fewer than 25% including information on sex, gender, race, or ethnicity, and fewer than 50% including age or country of origin. Such gaps in reporting make it difficult to determine who was represented in the datasets and how; additionally, for algorithms that are trained or tested using these datasets, their performance and safety across population subgroups cannot be evaluated.

Representative data are vital not only to assess the impact of diseases and related AI tools across different population subgroups, but also to evaluate the uptake of public health interventions. Leveraging nationally representative data, Sumali Bajaj and colleagues studied COVID-19 testing behaviours in England across different sociodemographic groups. They found that lateral flow device testing and reporting was lowest in the most deprived areas, despite infection prevalence being highest in these areas, and minoritised ethnic groups and those aged 75 years or older were less likely to use confirmatory PCR tests through most of the pandemic. These findings of systemic biases in the national COVID-19 testing programme could be used to design more equitable uptake of interventions in the future.

How can the representativeness of health datasets be increased? A key step is encouraging more diverse participation in research. Analyses of participants in COVID-19 clinical trials highlight inadequacies in the reporting and enrolment of diverse demographic groups. In September 2024, WHO published global guidance on best practices for clinical trials, calling on researchers to use more inclusive trial eligibility criteria so that under-represented populations can be involved, and on Member States and regulatory authorities to participate in constructive dialogue with these populations on identifying relevant research priorities and methodologies. International organisations and funders are recommended to synergise their resources and support and encourage open collaboration and data sharing.

In their Viewpoint, Sarah Jiang and colleagues pose that missing demographic information in a dataset is a larger concern than imbalanced but reported information, because encoded biases will not be as easy to detect. Adopting standards to ensure the proper documentation and use of health datasets, such as the forthcoming recommendations from the STANDING Together initiative, is thus imperative for transparent reporting of these datasets and to support researchers and AI developers in making more informed, responsible decisions regarding their use. Additionally, new methods such as synthetic data and oversampling of under-represented groups could aid the fair use of imbalanced datasets and help mitigate the impact of biases on health outcomes.

Open science practices can also be an effective means of addressing demographic gaps in datasets. Making datasets open access allows others to assess and determine the representativeness and appropriateness of the data for their purposes, and also enables more scrutiny of AI algorithms that are trained or tested using the data. But a robust governance model must be in place to ensure the data are used in a secure and ethical way.

Consistently and transparently reporting the representativeness of health datasets is pivotal to improve the quality of research evidence, and to bolster patient and public trust. Funders, regulators, and health policy makers should strongly encourage researchers and AI developers to utilise more inclusive study eligibility criteria, ensure the proper documentation, secure sharing, and use of health datasets, and perform relevant subgroup analyses to identify potential biases or performance variability across demographic groups. Collectively, the efforts outlined here should lead to more equitable datasets and innovations that benefit all. ■ *The Lancet Digital Health*

For more on **global COVID-19 statistics** see https://data.who.int/dashboards/covid19/deaths?n=o

For more on **CoViNet** see https://www.who.int/groups/who-coronavirus-network

For the **analysis by Joseph E Alderman and colleagues** see **Review** page e827

For the **study by Sumali Bajaj and colleagues** see **Articles** page e778

For more on the **representativeness of US COVID-19 trial participants** see *JAMA Intern Med* 2023; **183:** 50–60

For more on **WHO's guidance for best practices for clinical trials** see https://www.who.int/publications/i/item/9789240097711

For the **analysis by Sarah Jiang and colleagues** see **Viewpoint** page e871

For more on the **STANDING Together initiative** see https://www.datadiversity.org/home

For more on **open access data and governance** see *PLoS Digit Health* 2024; **3:** e0000599