Research and training recommendations for public health data science

The 2019 Next Generation Public Health meeting provided several useful recommendations on how big data and artificial intelligence (AI) could enhance public health.[1] To realise the full benefits of these developments, I propose two further recommendations. First, research studies should be done that will enable us to better understand the strengths, limitations, and applications of these new tools and data. Second, we need to train individuals who can bridge a skills gap that will enable the public health science community to fully engage with these developments.

Much of the historical research and investment into the use of big data and AI has focused on applications in genomics and personalised medicine. Today's public health challenges require evidence informed by complex systems models of the upstream drivers of health, such as the environment, education, and employment.[2] Multi-sectoral data about upstream health determinants can be linked to personal data (eg, from health-care records, mobile phones, and wearable devices) to increase the accuracy of complex systems models, our understanding of such systems, and opportunities for intervention. These data can also be used to inform the development and delivery of complex public health trials, natural experiments, and system change evaluations that have the potential to be undertaken more rapidly and efficiently than in the past. However, such data, tools, and methods must be evaluated and compared to more traditional study designs when applied to the public health data science tasks of description, prediction, causal inference,[3] and public health trials. These comparison studies should include cross-study and cross-workflow analyses with triangulation of results.[4] This practice will enable a better understanding of the relative strengths and weaknesses of different methods and facilitate decisions on when to use—and when to avoid—different study designs. Research, including modelling studies, should be done to help navigate the inherent tensions, and possible synergies, between interventions (including digital interventions) targeted at high-risk groups, and the application of universal approaches to tackle the health risks associated with the largest burdens of disease (including smoking, alcohol, diet, and physical activity).

To move forward with this work, public health training curriculums must evolve. Newly trained public health data scientists should be taught about how big data and AI can—and cannot—be applied to public health problems, and should be part of a transdisciplinary community that can collaborate within and beyond academia to influence the upstream drivers of health. Crucially, this training should highlight issues of health equity, which is at increased risk with the advent of algorithm-induced inequalities. For example, socially excluded populations are much less likely to be included in datasets used to train AI algorithms, resulting in these algorithms excluding and further marginalising these individuals. Additionally, the traditional skills of leadership, advocacy, and policy development will remain crucial if public health data science is to result in tangible improvements in the health of the public.

RWA is honorary consultant in public health data science at Public Health England and undertakes public health data science teaching and research.

Robert W Aldridge
r.aldridge@ucl.ac.uk

Institute of Health Informatics, University College London, London NW1 2DA, UK

1 The Lancet Public Health. Next generation public health: towards precision and fairness. *Lancet Public Health* 2019; **4:** e209.
2 Rutter H, Savona N, Glonti K, et al. The need for a complex systems model of evidence for public health. *Lancet* 2017; **390:** 2602–04.
3 Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance* 2019; **32:** 42–49.
4 Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol* 2016; **45:** 1866–86.