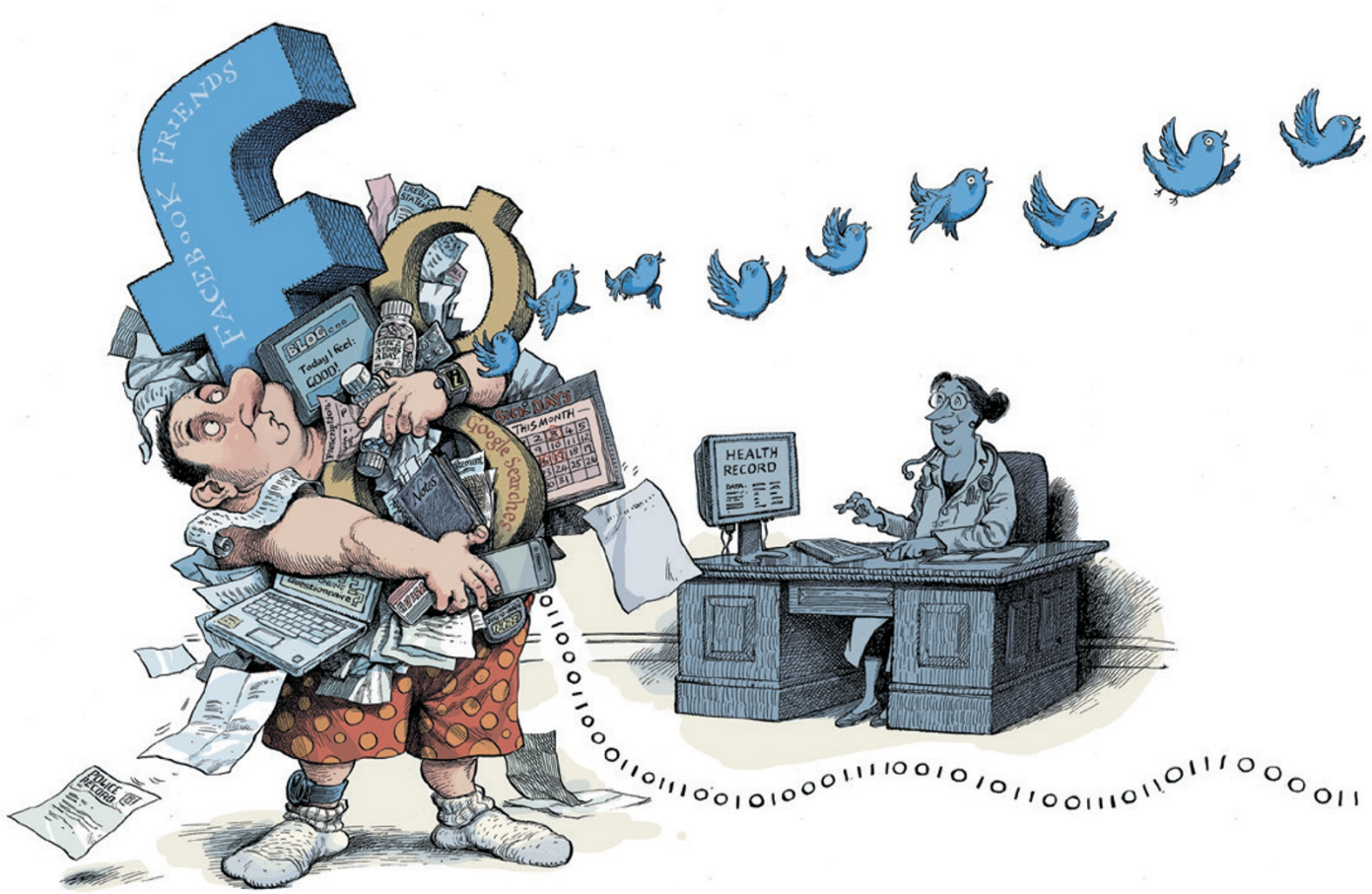
COMMENT

DATA A call for open and democratic information aggregators and filters **p.33**

NEUROSCIENCE Pseudoscientific justifications for torture are roundly debunked **p.35**

FILM Steve Jobs biopic more sketch than complex study **p.36**

EMISSIONS Transport lessons for COP21 from Volkswagen scandal **p.38**



ILLUSTRATIONS BY DAVID PARKINS

Make sense of health data

Develop the science of data synthesis to join up the myriad varieties of health information, insist **Julian H. Elliott**, **Jeremy Grimshaw** and colleagues.

If you are wondering whether exposure to some chemical could increase your chances of getting colon cancer, you could easily find supportive evidence from animal experiments. You might then discover that epidemiological studies tell a different story.

There have never been more options when it comes to measuring factors relevant to health. We can sequence our entire genomes and those of our bacteria, viruses and tumours. In principle, every visit to the doctor can be tracked from electronic medical records. Information on physiology, behaviours, diets, movements and interactions with others can be extracted from wearable devices, smartphone apps and social-networking sites[1]. And thanks to the open-access movement and a shift in data-sharing norms, more data are being made publicly available.

Yet sifting through the information to find answers to questions about health is becoming increasingly difficult, even for the experts. The data exist in disparate domains, are generated using different methods, and are stored in different infrastructures — from the private servers of hospitals to global platforms, such as dbGaP, an open database of genotypes and clinical information.

POOLING DATA

We believe that to consolidate data from different sources into comprehensive and coherent bodies of evidence on which decision-makers can act, researchers need to better exploit current methods and tools for data synthesis — and to develop superior ones.

Researchers usually try to obtain insights by pooling the same kind of data, such as ▶

from clinical trials. But because different study and data types tend to have distinct strengths and weaknesses, a much richer understanding can emerge when different kinds of information are combined.

The drug cisapride, for instance, was licensed in the United States in 1993 to treat heartburn, on the basis of data collected in clinical trials over ten years. Yet the drug's association with fatal heart-rhythm disturbances[2] was understood only when data from clinical trials were consolidated with those from large, long-term cohort studies, which recorded cisapride's effects in thousands of people.

Likewise, the picture obtained from conventional influenza surveillance (which involves collecting data from primary-care clinics) can lag behind what is actually happening on the ground. Google collects real-time information based on the use of search terms related to flu symptoms, but these findings can be inaccurate. The best insights almost certainly come from aggregating these different data types[3].

So how can we bring together the multiple, extremely diverse data sets that are now becoming available?

Formal methods for 'evidence synthesis' — in which multiple sources of data are combined to obtain new insights — were first developed in the social sciences in the 1970s. The techniques have since been adapted in many branches of science, and they underpin high-impact decision-making, for example in drug licensing[4]. They generally involve identifying and collating all the available and relevant data; assessing each data source's strengths and vulnerability to bias; and deciding how to handle the different sources of data depending on their rigour and the question being asked (some data may be excluded, for instance). Then, if appropriate, a meta-analysis or qualitative assessment can be conducted, incorporating the information[5].

For example, a UK group combined[6] data from clinical trials with those from cohort studies in a meta-analysis to assess the effectiveness of anti-D, a drug given to some pregnant women to prevent them from producing antibodies against their babies. In this case, potential sources of bias, such as different clinics providing care for the women in cohort studies, were systematically identified, and their impact was minimized.

Yet many researchers immersed in the combination and analysis of large data sets that are vulnerable to spurious correlations, such as genomic or electronic-medical-record data, are unaware of evidence-synthesis tools and their potential usefulness. Conversely, many experts in evidence synthesis are unfamiliar with the methods often used to analyse large data sets relevant to health.

We believe that the core elements of evidence synthesis must be combined with other data sciences to develop new ways to make sense of diverse data.

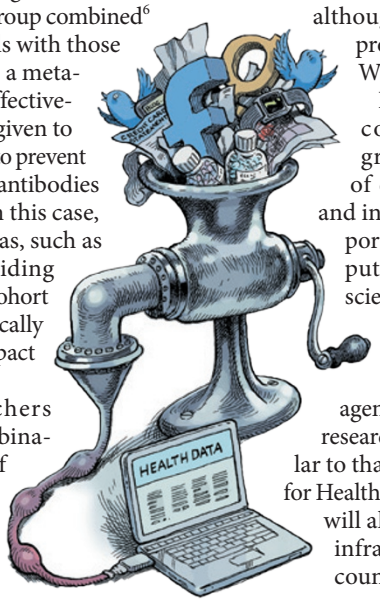
MANAGING BIAS

Scientists need to work out why, when and how to combine diverse data — for instance, should physical-activity data from clinical records, online questionnaires and wearable devices be combined? As well as addressing when and how to combine diverse individual-level data, scientists need to grasp the risks of bias associated with each data type and incorporate such risks into their analyses. For clinical trials and observational studies of the effects of interventions, analysts can use the Cochrane Risk of Bias approach. Similar methods are needed to enable the detection and reduction of bias in other data types, such as social-networking and mobile-phone data.

"Society does not need more islands of data analysis."

Also needed are agreed ways to capture and represent information on potential sources of bias. Organizations investing in infrastructure and standards for health data, such as Health-Level 7, need to incorporate this layer of metadata (data about data) into their systems.

Methods to deal with bias must be incorporated into new analytical systems developed to guide decision-making in health care — including those based on natural-language processing and machine learning. Transparent and independent evaluations of these new systems will also be important, although challenging to achieve for proprietary systems such as IBM Watson.



In the short to medium term, conferences, funding programmes and a restructuring of departments in universities and institutes will be crucial to support collaborations between computational biologists, computer scientists, clinical and population-health researchers and specialists in evidence synthesis. For instance, major granting agencies should invest in dedicated research-methods programmes similar to that of the UK National Institute for Health Research. Targeted investment will also be needed to develop data infrastructure in poor regions and countries. In the long term, a new type of analyst, adept at appraising and combining diverse data types appropriately, may emerge.

JOINING THE DOTS

What could these shifts mean in practice? One of the aims of the US Precision Medicine Initiative (PMI) is to prevent people from getting cancer. This means understanding the effects of myriad genomic, behavioural and environmental factors and their interactions. The value of the initiative will be enhanced if data from these very different domains can be combined appropriately and easily.

Another aim of the initiative is to develop new cancer therapies. Better systems for data synthesis would inform drug development with richer and more accurate insights from the 'omics' sciences, animal studies and early human trials. Moreover, health-care funders such as Britain's National Health Service and Medicare in the United States could better understand a drug's benefits and harms in the real world by synthesizing data from clinical trials, cohort studies, patient experiences reported through mobile and social applications, and drug-surveillance systems. (These include the US Sentinel Initiative and the Canadian Network for Observational Drug Effect Studies, which pool data from different health-care systems to monitor the adverse effects of licensed drugs.)

We are not proposing a one-model-fits-all approach. But society does not need more islands of data analysis that support conflicting inferences. As large and diverse data sets become ever more plentiful, we must ensure that rigorous and trustworthy methods to make sense of the data are developed in parallel. ■

Julian H. Elliott *is senior research fellow at the Australasian Cochrane Centre at Monash University, and head of clinical research in the Infectious Diseases Unit at Alfred Hospital, Melbourne, Australia.* **Jeremy Grimshaw** *is senior scientist at Ottawa Hospital Research Institute and professor of medicine at the University of Ottawa, Canada.* **Russ Altman**, **Lisa Bero**, **Steven N. Goodman**, **David Henry**, **Malcolm Macleod**, **David Tovey**, **Peter Tugwell**, **Howard White**, **Ida Sim**.
e-mail: julian.elliott@alfred.org.au

1. Weber, G. M., Mandl, K. D. & Kohane, I. S. *J. Am. Med. Assoc.* **311,** 2479–2480 (2014).
2. Wysowski, D. K. & Bacsanyi, J. *N. Engl. J. Med.* **335,** 290–291 (1996).
3. Lazer, D., Kennedy, R., King, G. & Vespignani, A. *Science* **343,** 1203–1205 (2014).
4. Institute of Medicine. *Finding What Works in Health Care: Standards for Systematic Reviews* (National Academies Press, 2011).
5. Chalmers, I. *Ann. Am. Acad. Pol. Soc. Sci.* **589,** 22–40 (2003).
6. Turner, R. M. *et al. PLoS ONE* **7,** e30711 (2012).

The authors declare competing financial interests. For details, and for full author affiliations, see go.nature.com/scxwp9.