MEDICINE

Big data meets public health

Human well-being could benefit from large-scale data if large-scale noise is minimized

By **Muin J. Khoury**[1,2] *and* **John P. A. Ioannidis**[3]

From validity to utility. Big Data can improve tracking and response to infectious disease outbreaks, discovery of early warning signals of disease, and development of diagnostic tests and therapeutics.

In 1854, as cholera swept through London, John Snow, the father of modern epidemiology, painstakingly recorded the locations of affected homes. After long, laborious work, he implicated the Broad Street water pump as the source of the outbreak, even without knowing that a *Vibrio* organism caused cholera. "Today, Snow might have crunched Global Positioning System information and disease prevalence data, solving the problem within hours" (*1*). That is the potential impact of "Big Data" on the public's health. But the promise of Big Data is also accompanied by claims that "the scientific method itself is becoming obsolete" (*2*), as next-generation computers, such as IBM's Watson (*3*), sift through the digital world to provide predictive models based on massive information. Separating the true signal from the gigantic amount of noise is neither easy nor straightforward, but it is a challenge that must be tackled if information is ever to be translated into societal well-being.

The term "Big Data" refers to volumes of large, complex, linkable information (*4*). Beyond genomics and other "omic" fields, Big Data includes medical, environmental, financial, geographic, and social media information. Most of this digital information was unavailable a decade ago. This swell of data will continue to grow, stoked by sources that are currently unimaginable. Big Data stands to improve health by providing insights into the causes and outcomes of disease, better drug targets for precision medicine, and enhanced disease prediction and prevention. Moreover, citizen-scientists will increasingly use this information to promote their own health and wellness. Big Data can improve our understanding of health behaviors (smoking, drinking, etc.) and accelerate the knowledge-to-diffusion cycle (*5*).

But "Big Error" can plague Big Data. In 2013, when influenza hit the United States hard and early, analysis of flu-related Internet searches drastically overestimated peak flu levels (*6*) relative to those determined by traditional public health surveillance. Even more problematic is the potential for many false alarms triggered by large-scale examination of putative associations with disease outcomes. Paradoxically, the proportion of false alarms among all proposed "findings" may increase when one can measure more things (*7*). Spurious correlations and ecological fallacies may multiply. There are numerous such examples (*8*), such as "honey-producing bee colonies inversely correlate with juvenile arrests for marijuana."

The field of genomics has addressed this problem of signal and noise by requiring replication of study findings and by asking for much stronger signals in terms of statistical significance. This requires the use of collaborative large-scale epidemiologic studies. For nongenomic associations, false alarms due to confounding variables or other biases are possible even with very large-scale studies, extensive replication, and very strong signals (*9*). Big Data's strength is in finding associations, not in showing whether these associations have meaning. Finding a signal is only the first step.

Even John Snow needed to start with a plausible hypothesis to know where to look, i.e., choose what data to examine. If all he had was massive amounts of data, he might well have ended up with a correlation as spurious as the honey bee–marijuana connection. Crucially, Snow "did the experiment." He removed the handle from the water pump and dramatically reduced the spread of cholera, thus moving from correlation to causation and effective intervention.

How can we improve the potential for Big Data to improve health and prevent disease? One priority is that a stronger epidemiological foundation is needed. Big Data analysis is currently largely based on convenient samples of people or information available on the Internet. When associations are probed between perfectly measured data (e.g., a genome sequence) and poorly measured data (e.g., administrative claims health data), research accuracy is dictated by the weakest link. Big Data are observational in nature and are fraught with many biases such as selection, confounding variables, and lack of generalizability. Big Data analysis may be embedded in epidemiologically well-characterized and representative populations. This epidemiologic approach has served the genomics community well (*10*) and can be extended to other types of Big Data.

There also must be a means to integrate knowledge that is based on a highly iterative process of interpreting what we know and don't know from within and across scientific disciplines. This requires knowledge management, knowledge synthesis, and knowledge translation (*11*). Curation can be aided by machine learning algorithms. An example is the ClinGen project (*12*) that will create centralized resources of clinically annotated genes to improve interpretation of genomic variation and optimize the use of genomics in practice. And new funding, such as the Biomedical Data to Knowledge awards of the U.S. National Institutes of Health, will develop new tools and training in this arena.

[1]*Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA.* [2]*Epidemiology and Genomics Research Program, National Cancer Institute, Bethesda, MD 20850, USA.* [3]*Stanford Prevention Research Center and Meta-Research Innovation Center at Stanford, Stanford University, Palo Alto, CA 94305, USA. E-mail: muk1@cdc.gov; jioannid@stanford.edu*

ILLUSTRATION: V. ALTOUNIAN/*SCIENCE*

Another important issue to address is that Big Data is a hypothesis-generating machine, but even after robust associations are established, evidence of health-related utility (i.e., assessing balance of health benefits versus harms) is still needed. Documenting the utility of genomics and Big Data information will necessitate the use of randomized clinical trials and other experimental designs (*13*). Emerging treatments based on Big Data signals need to be tested in intervention studies. Predictive tools also should be tested. In other words, we should embrace (and not run away from) principles of evidence-based medicine. We need to move from clinical validity (confirming robust relationships between Big Data and disease) to clinical utility (answering the "who cares?" health impact questions).

As with genomics, an expanded translational research agenda (*14*) for Big Data is needed that goes beyond an initial research discovery. In genomics, most published research consists of either basic scientific discoveries or preclinical research designed to develop health-related tests and interventions. What happens after that in the bench-to-bedside journey is a "road less traveled" with <1% of published research (*15*) dealing with validation, evaluation, implementation, policy, communication, and outcome research in the real world. Reaping the benefits of Big Data requires a "Big Picture" view.

Bringing Big Data to bear on public health is where the rubber meets the road. The combination of a strong epidemiologic foundation, robust knowledge integration, principles of evidence-based medicine, and an expanded translation research agenda can put Big Data on the right course. ■

REFERENCES

1. Harvard School of Public Health (2014); www.hsph.harvard.edu/news/magazine/big-datas-big-visionary.
2. A. Standen, KQED Science (2014); blogs.kqed.org/science/audio/how-big-data-is-changing-medicine.
3. G. Eysenbach, *Am. J. Prev. Med.* **40** (suppl. 2), S154 (2011).
4. National Institutes of Health, BD2K (2014); bd2k.nih.gov/index.html#sthash.0uOeCsq3.dpbs.
5. R. High, J. Low, *Scientific American* blogs (2014); blogs.scientificamerican.com/mind-guest-blog/2014/10/20/expert-cancer-care-may-soon-be-everywhere-thanks-to-watson.
6. D. Butler, *Nature News* (2013); www.nature.com/news/when-google-got-flu-wrong-1.12413.
7. J. P. A. Ioannidis *et al.*, *PLOS Med.* **2**, e24 (2005).
8. Spurious Correlations (2014); tylervigen.com.
9. J. P. A. Ioannidis, E. Y. Loy, R. Poulton, K. S. Chia, *Sci. Transl. Med.* **1**, 7ps8 (2009).
10. M. J. Khoury, M. Gwinn, M. Clyne, W. Yu, *Genet. Epidemiol.* **35**, 845 (2011).
11. M. J. Khoury *et al.*, *Genet. Med.* **14**, 643 (2012).
12. National Human Genome Research Institute (2013); www.nih.gov/news/health/sep2013/nhgri-25.htm.
13. J. P. A. Ioannidis, M. J. Khoury, *Genome Med.* **5**, 32 (2013).
14. S. D. Schully, M. J. Khoury, *Appl. Transl. Genomics* www.sciencedirect.com/science/article/pii/S2212066114000313 (2014).
15. M. Clyne *et al.*, *Genet. Med.* **16**, 535 (2014).

10.1126/science.aaa2709

PHOTOS: JANICE WIESE-FALES, FOAMS PREPARED BY HARITH H. AL-MOAMERI

MATERIALS DESIGN

Simulation as a tool for biopolymers design

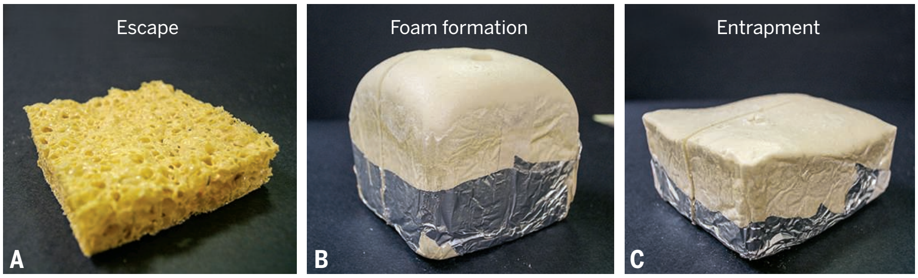
New materials made from biobased raw materials can be formulated faster with computer simulation of their reaction chemistry

By **Galen J. Suppes**

Many polymers are made from simple small molecules (monomers) derived through many processing steps from petroleum. Biologically sourced materials are more sustainable, but to be cost competitive, new products will likely need to be based on different structures composed of natural oils or alcohols that have undergone minimal processing. For example, plant-derived oils can be made into polymers, but these oils can bear many more functional groups in variable positions, which poses considerable challenges to controlling their reaction chemistry. The use of computer simulation based on experimental inputs offers a way to control this complexity and transform materials design, just as process simulators have transformed engineering design. This perspective presents a case that simulation is ready to change the way we research, develop, and design thermoset polymer recipes (formulations).

THERMOSET POLYMERS. The spray foam insulation dispensed from pressurized cans for filling cracks and crevices is an example of a thermoset polymer. The two components, isocyanate and polyol monomers, combine during the spray and react to form a polyurethane; the degree of foaming is controlled by in situ gas generation from reactions (e.g., water and isocyanate) or evaporation. Unlike thermoplastics, which can be heated and then molded, thermosets adopt a final fixed shape.

The degrees of freedom of a typical urethane foam formulation include selecting types and amounts of monomers (4 to 7 degrees of freedom), catalysts (2 to 4), surfactants (2), and blowing agents (2 to 4). Biobased materials (*1*–*5*), fillers, and fire retardants can add additional degrees of freedom. Historically, the handling of these 10 to 23 degrees of freedom in specifying a urethane has been much like an "art" in the hands of a Ph.D. chemist with decades of experience. Predicting the reaction progress is not simple because of the many simultaneous physical and chemical processes. Hundreds of customer-specific urethane formulations depend on the proper combinations of these parameters and must meet constraints on product performance, raw material costs, and throughput (i.e., cure time).



Just right. Curing a polyurethane foam is like baking bread—it must trap gas bubbles in order to rise (**B**). If the "dough" is too runny (too low in viscosity), bubbles escape (**A**); too stiff (too high in viscosity), the gas is trapped and no bubbles form (**C**).

USE OF SIMULATION. Chemical process simulators have transformed the design of many reaction and separation processes when applications have a large numbers of degrees of freedom, simultaneous solution of multiple governing equations, and large markets that provide good returns on investment in simulation process. Simulation is used for both designing new processes and controlling existing processes. With cumulative markets exceeding $20 billion

Department of Chemical Engineering, University of Missouri, Columbia, MO, USA. E-mail: suppesg@missouri.edu