Big data and the future

At the beginning of her career **Sherri Rose** discusses big data and stands amazed at its potential.

The Human Genome Project began in 1990. By 2003 it had produced a map of the approximately 20 000–25 000 genes that make the human species what it is. It was a tremendous effort, a defining project in humankind's understanding of itself. The collection, analysis, and interpretation of that huge amount of data took 12 years of concentrated effort by the best human minds and the best human-made machines.

The human mind has not changed since then, but the human-made machines have. So has what they can do. The first "big data" research area I was exposed to as a young undergraduate student was genomics. Steven Salzberg, Professor of Medicine and Biostatistics at Johns Hopkins University, summarised today's genomics landscape for me: "Next-generation sequencing technology can now generate more data in a single day than the entire Human Genome Project generated in 12 years. It has transformed biomedical science." It was apparent to us as novices that working as a statistician in genomics would require more than a deep understanding of the biology involved; it would need also a special arsenal of statistical, computational, and technological tools. There is just too much data for the old ways to handle.

© iStockphoto.com/Pgiam

Even just looking at the data brings problems. "Simply moving this data around presents major challenges to many scientists and institutions: their networks just aren't fast enough", he says. And that is before you start working on it: "Analysing the data is a much bigger problem. With such large data sets, it is all too easy to find rare statistical anomalies and to confuse them with real phenomena." When there are millions of data points, and many tests, false positives are much more likely.

With such large data sets, it is all too easy to find rare statistical anomalies and to confuse them with real phenomena

"Big data" has become a hot topic across many disparate fields, and for good reason. One burgeoning new area that generates very large data sets is the study of brain images. Ani Eloyan, a postdoctoral fellow at Johns Hopkins Bloomberg School of Public Health, is part of the team that won a recent prediction contest examining attention deficit hyperactivity disorder (ADHD), the 2011 ADHD-200 Global Competition. They used neuroimaging data and other information to categorise subjects into neurotypical, ADHD primary inattentive type, or

ADHD combined type diagnoses. Understanding the unique complexities of imaging data is no small task. "Brain imaging data mostly consist of collections of three-dimensional arrays (of, for example, intensities) collected over time resulting in a four-dimensional array for each subject", she says (for a related example, see Feigelson and Babu's image from astronomy on page 23). "The first major issue in analysing these data is the simple fact that our brains are very different in size, shape and so on. In many cases the transformation of the matrices into a common space – a form in which they can be compared to each other – is still an open problem which is hindering the analysis of the data."

The analysis of large comprehensive medical databases is another area where statisticians are lending their skills. These databases are even making appearances in the mainstream media, thanks to a new competition. Following in the footsteps of the $1 million Netflix Prize (see page 40), where teams developed algorithms to improve upon the content provider's existing recommendation system for movies, is an even larger big-data prize that is perhaps more socially useful: the $3 million Heritage Health Prize Competition. Its goal is to predict future hospitalisations using existing high-dimensional patient data. A behemoth example of a massive clinical database is the US Food and Drug Administration's Sentinel Initiative, which aims to monitor drugs and medical devices for safety over time. The end result of this programme will be a national electronic system, and the new system already has access to 100 million people and their medical records. Consider the volume of medical data that one person can accumulate over a very few years: repeated measurements of blood pressure, lung function, antibody concentrations, digitalised X-rays and scans and the rest. Multiply that by 100 million and you get an idea of the size of the database.

Imagine you have in your hands the ability to understand the yin and yang between tolerance and immune response, the power to decipher the secrets of complex diseases

As one can imagine, the sheer scale of this project and its longitudinal nature – its following of patients through time – provide interesting statistical challenges. One complexity is accurately defining the data. In this case, among other considerations, it involves acknowledging that subjects "drop out" and are not followed for the entire time period scientists are studying. The model has to include a mechanism that generates these missing values, as subjects do not drop out at random: issues such as drug toxicity could be leading to drop out – which is a bit of a problem if drug toxicity is the very thing you are trying to study. The traditional assumptions of parametric modelling are not likely to be supported by what is known about how the data was generated. Mark van der Laan, Professor of Biostatistics and Statistics at the University of California, Berkeley, and authority in statistical learning with missing data, is working on the Sentinel Initiative. "We need to use the state of the art in estimation without relying on restrictive assumptions; we need methods that aim to learn from these large data sets as much as the data allow."

Electronic medical records are only part of big data. They are being linked with other big data sets to study, for example, environmental issues such as air quality. Examining the health effects of air quality brings in the additional component of geography: different regions have different particles in the air. Cory Zigler is a postdoctoral fellow at Harvard School of Public Health. He has recently investigated the causal effects of the 1990 Clean Air Act amendments on millions of Medicare beneficiaries, with applications to health and environmental policy. Dr. Zigler illuminated the many issues environmental biostatisticians face:

> There are satellites measuring markers of ambient air quality at increasingly fine spatial and temporal resolutions. Couple that with the wealth of health data being collected in administrative databases, and we're faced with a big data challenge in environmental epidemiology. But all the data in the world won't change some of the salient issues such as the fact that people who live near one another share many things in common in addition to the air they breathe. Teasing out the health effects of air pollution from other factors requires thoughtful statistical reasoning throughout the entire process: you must define the right question, choose the right spatial and temporal resolution of the data, ultimately apply the right analytical methods and interpret them correctly. This must be a combined effort from people with a wide array of quantitative skills.

Disentangling the impact of the community, neighbourhood, and household effects is a far-reaching challenge found in many other fields as well.

What will the big data allow, once we have learned how to handle it properly? Alessio Fasano, Director of the Center for Celiac Research at the University of Maryland School of Medicine, is an expert on gluten-related disorders. His vision of the future is inspiring:

> Imagine that you have in your hands the ability to unveil the secrets of human biology, to establish how the human host interacts and communicates with the "parallel civilisation" of bacteria living in symbiosis with us, to understand the yin and yang between tolerance and immune response, and the ability to turn on and off autoimmune diseases at will. Imagine, in other words, that you have the power to decipher the secrets of complex diseases, so that innovative preventive and therapeutic interventions can be developed. All this is theoretically possible with celiac disease, the only autoimmune disease for which the environmental trigger is known. However, these goals are achievable only if robust statistical methodologies are applied to elaborate the enormous amount of data that we have recently acquired, thanks to advances in our knowledge about celiac disease pathogenesis. Trying to make sense of the complexity of celiac disease without fundamentals in statistics is like trying to decipher Egyptian hieroglyphics without having the key to interpret them.

Dr. Fasano is leading innovative new projects studying the introduction of gluten in infants and their microbial environment, science that will advance human knowledge radically, but only with the collaboration of medical, scientific – and statistical – experts.

As a recent statistical trainee, a freshly-minted (bio)statistics PhD, I believe we as young researchers have a responsibility. Our new projects, will be interdisciplinary. We must ground them in the fundamentals and principles of statistics. They will involve many new colleagues and disciplines, and will increasingly be our future. Small wonder that young researchers are getting excited about big data.

Sherri Rose is an NSF postdoctoral fellow in biostatistics at the Johns Hopkins Bloomberg School of Public Health. She recently coauthored the book *Targeted Learning: Causal Inference for Observational and Experimental Data* for the Springer Series in Statistics.