Using Big Data to Solve Autism and Other Mysteries

Recent newspaper articles have highlighted autism studies that lean toward genetic causes on the one hand and environmental on the other. One notes correlations with the age of fathers and the genetic mutations that we all inherit but that increase with a father's age. Another suggests that we have a weakened resistance to germs because we aren't exposed to as many in our cleaner, less outdoor society. Most of us are also familiar with past studies that failed to find evidence for the popular thesis that immunizations given to young children increase the probability of autism.

These autism studies are mere examples of the many types of epidemiological research that try to investigate outbreaks of disease, assess exposure risks, or figure out why certain populations seem to be more or less immune to various health threats. The research often looks for both good and bad exceptions to averages. Malcolm Gladwell's introduction to his popular book *Outliers*, for instance, points out the "Roseto Mystery": the studies by Stewart Wolf and John Bruhn on why people living in Roseto, Pennsylvania, have relatively fewer heart attacks and live longer than those living elsewhere.

Progress? Yes. Yet someday these one-off studies will be likened to the late Middle Ages when it comes to medical science research. Not for their conclusions, but for the years, and, in some cases, decades of ex post data gathering required before any conclusions are reached.

Imagine a different world, in which data on these populations had already been transferred through electronic health records to the Centers for Disease Control (CDC) or a similar agency. In the world of big data, one doesn't always work from casual observation to hypothesis to painstaking data gathering—sometimes guessing at the right sample populations to begin following, perhaps for years and decades into the future. In this imagined world, much data on them and on many comparison populations would already have been gathered.

Research, of course, is always somewhat haphazard. You never know what you are going to find, and when you find it, you need to determine whether "it" is genuine or an anomaly. But with large amounts of data already available, the odds of finding "it" and proving "it" are magnified.

In this new world, research could also proceed from computer-generated detections of correlations to hypothesis and theory, rather than the other way around—in some ways reversing the traditional methodology of modern science from Descartes onward. Thus, correlations at times are found even when not originally hypothesized, and discoveries may abound. Although some relationships may simply reflect random chance—flip a coin enough times and heads will eventually pop up 20 times in a row—rechecking is easy by testing different subsets of big data sets.

With so many relationships to be examined, whether with traditional or new methodology, new understandings can proliferate, as well as quicker rejection of hypotheses that cannot be substantiated. For autism, for instance, we would know much more quickly about its prevalence in different geographic regions with different environmental exposures and about the effectiveness of various interventions, from diets to drugs to early educational efforts.

Similarly, we would uncover much earlier warning signals, whether of a sudden flu epidemic or an increase in the prevalence of Alzheimer's or heart disease by region, sex, race, or other characteristic.

For several years I was privileged to work with a group of very fine doctors, researchers, lawyers, economists, and other health care experts on the National Committee for Vital and Health Statistics. Its primary interest then—and, to some extent, now—was to expand the use of electronic health records (EHRs).

Many associate electronic health records with better transmission of information from one hospital, doctor, or other health care provider to the next. After Katrina, for instance, we were all appalled at the inability of victims to have their medical records available to those treating them in neighboring jurisdictions.

Others recognize that EHRs make it easier to detect sources of individual health problems. Thanks to EHRs, most pharmacists now get computer-generated information on drugs that contravene each other; doctors can plug symptoms into computers that spew out lists of possible causes, including some they might have neglected, forgotten, or never learned.
But, for many of us on the committee, we ultimately hoped to create a world in which much faster, more thorough, and more comprehensive public health research could be performed on the causes and possible cures for disease, malignancies, and chronic health conditions, outbreaks of new health problems, and local or regional stories of failure or success in places like Roseto. How many, when reading a story about a place like Roseto, realize that in today's world we shouldn't have to wait decades to accidentally discover such geographical variations?

What blocks progress? Many forces. For one, lack of incentives for doctors. Mine have complained about the extra workload and the increased ability of insurance companies to sue them. To try to overcome this, every recent president has increased exponentially the subsidies available for adopting EHRs. For another, standards. Not just for nomenclature but for software that will handle data and make it transferable. Privacy is a third major obstacle. Many individuals don't want their records sent to a centralized location, as security can never be 100 percent.

In a talk I gave several years ago at the National Academies, I argued that we may achieve real progress only when consumers begin to demand these improvements. What if a subset of parents of autistic children demanded that their children's health records be gathered together at the CDC or some other place? They would work with IT professionals, medical researchers, doctors, and teachers with special knowledge of autism to create common data fields. With enough participants, data provided by only a subset of cases would be sufficient for some research.

Add to these parents of autistic children the children of parents with Alzheimer’s or simply people like me who know the auto-immune problems that my children could have inherited from both sides of the family. What if we were to rank our doctors and their practices by how well they participate in such shared data gathering? What if some foundations helped organize these consumers?

In the end, organizing consumers so they can demand the possible may be more important than all the money in the world, which is what we seem to be spending on health care already without the progress we can and should be making.

We are on the cusp of great possibilities in health research, a scientific revolution of sorts. Big data, electronic health records, and government committees provide some of the wherewithal, but we've got to make the leap.