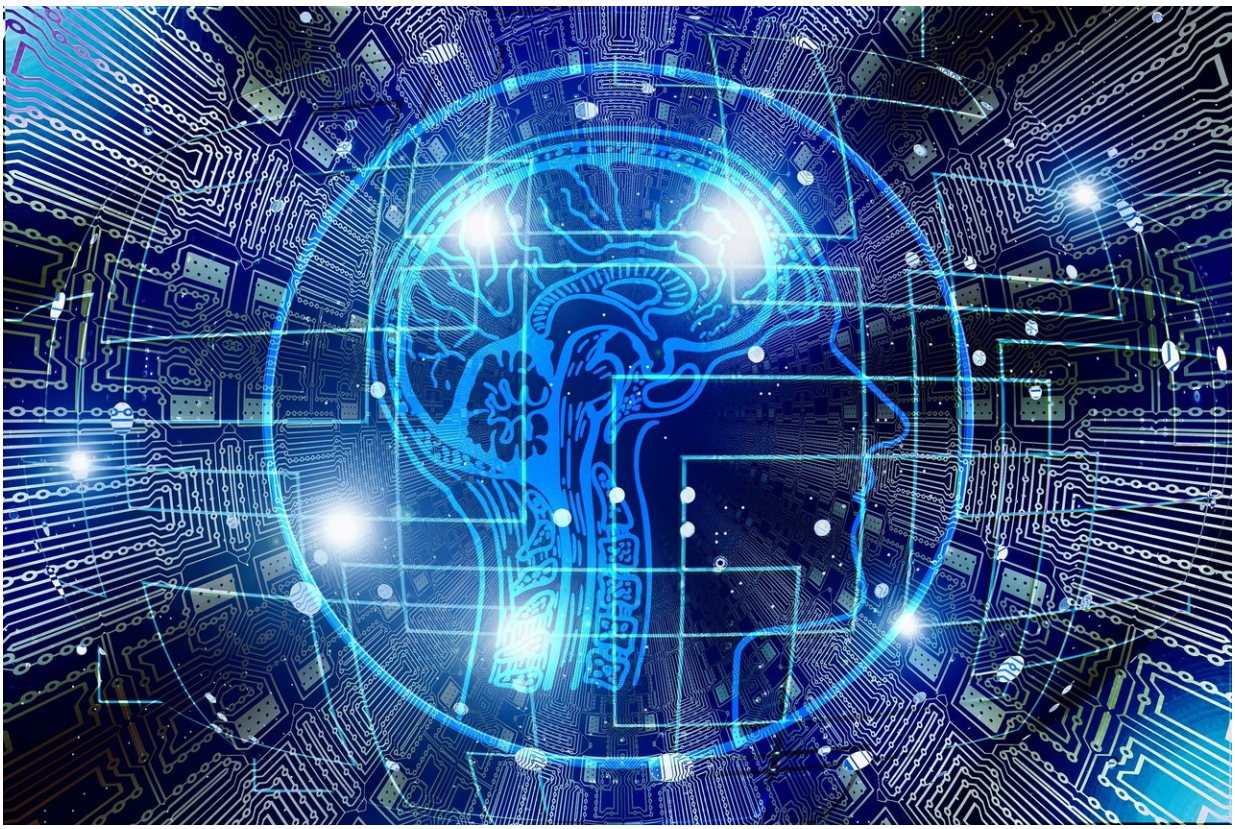
The surprising power of small data—more information isn't necessarily better in health care or business

November 22 2018, by Lee Simmons



Credit: CC0 Public Domain

Chronic conditions like heart disease and diabetes have been on the rise for decades. They're the number one cause of death and disability in the

U.S. today and one reason why health care costs are out of control.

So identifying people at risk for chronic conditions before they get sick makes a lot of sense. At the very least, early intervention can often slow the pace of disease and improve patients' quality of life—and in doing so, potentially save billions of dollars in medical costs.

That's why many employers—some 50%, according to a RAND report—sponsor incentivized wellness programs for their workers. Along with gym discounts, these programs typically include a health-risk assessment in the form of lab tests used to calculate each person's risk factors for common diseases. Those at risk are then offered extra preventive care and oversight.

Unfortunately, the expected benefits don't always materialize, says Mohsen Bayati, an associate professor of operations, information, and technology at Stanford Graduate School of Business. Several studies have found that such programs can end up costing more money than they save. One likely reason, he says, is that the risk assessments themselves aren't all that accurate.

"If you wrongly identify someone as high risk—a so-called 'false positive'—you pay for unnecessary services," Bayati says. "And if you miss someone who truly is at risk—a false negative—then you still get hit with those huge medical bills in the future."

One solution, he says, would be to run a more elaborate panel of tests. But that would also increase cost. "Lab tests are expensive. Companies are doing this for lots of employees, so they look at a fairly small set of standard biomarkers. And then the detection ability isn't very strong."

Instead, Bayati says, the key to making these preventive programs work is to improve the selection of biomarkers. But how do you do that? To

put it more rigorously: How do you choose a minimal set of markers that will maximize the diagnostic power over a range of diseases?

That's the puzzle Bayati tackled in a recent paper, which he coauthored with two Stanford colleagues: Sonia Bhaskar, Ph.D., a former Stanford research assistant who now works as a data scientist at Netflix, and Andrea Montanari, a statistics and electrical engineering professor. Using some technical jujitsu from the field of machine learning, they developed a method that can be used for any group of target diseases or program budget level.

When they tested it on medical records for some 75,000 patients, they found that it could predict a group of nine serious diseases with unexpected accuracy. "We were surprised," Bayati says. Compared with a hypothetical Cadillac-care assessment with no limit on the number of biomarkers, theirs would cost far less, yet have almost the same level of predictive power.

And maybe there's a general lesson here, in this era of Big Data. "You have to wonder," Bayati muses. "In every industry, companies are investing resources to gather more and more data. We're putting sensors on everything, just because we can, and frankly, it isn't all necessary or useful."

Too Much Information

Traditionally, health-risk assessments have been designed by figuring out the best markers for each disease in isolation and adding them to a list. "Hospitals are getting more sophisticated in how they identify biomarkers, with advanced statistics and now AI," Bayati says. "But it's all done one disease at a time."

You could potentially build an effective test panel this way, he says, but

it would require far too many biomarkers. So in practice, compromises are made and accuracy declines. Instead, Bayati and his colleagues added a second step to the analysis: "We said, let's start with that complete list and then see if we can simplify it in a better way to minimize the loss of diagnostic power."

To do that, they drew on some techniques from high-dimensional statistics that are used in machine learning. "The fundamental question is, if you have too much information, how can you narrow it down to the most useful smaller set of information? How do you reduce the dimensions of the data set?"

The math is involved, but basically, the key to solving that "TMI" problem is to jointly optimize the selection of biomarkers. Instead of finding the best ones for each disease separately, decide first how many biomarkers you want—the researchers settled on 30—and then maximize the predictive power, over all possible combinations, for the whole set of diseases at once.

The model works because many biomarkers signal more than one disease. High blood glucose, for instance, may be a sign of diabetes, but also kidney disease, liver disease, or heart disease. Abnormal levels of alkaline phosphatase are associated with heart disease, liver disease, and cancer. "If your selection process doesn't take these overlaps into account, you're throwing away information," Bayati says.

No Limit to Objectives

The power of the method Bayati and his colleagues outline is that it can be used to pursue multiple goals at once. What's more important in health-risk assessments: accuracy or cost? Both, of course. Do we want to predict Alzheimer's or arterial disease? Yes.

"There's no limit to the number of goals," Bayati says. "You could list 20, 30, 100 objectives that you want to optimize over. And then you can narrow down the information you need to collect—because at some point, adding objectives doesn't require additional data."

If it helps to fulfill the promise of corporate wellness programs, that's a big deal for health care. But this approach can also be used to improve a range of business and public policy operations. What's crucial, Bayati says, is to be clear on the objectives. Computers can do the analysis, but humans have to tell them what to optimize.

And that's a step, he thinks, companies too often gloss over. "Sometimes it seems that firms are just rushing to accumulate data and asking questions later. But more information isn't necessarily better. What matters is knowing what to look at. Our paper is a step in that direction."

Provided by Stanford University

Citation: The surprising power of small data—more information isn't necessarily better in health care or business (2018, November 22) retrieved 3 May 2024 from https://phys.org/news/2018-11-power-small-datamore-isnt-necessarily.html