Data mining for big data from Electronic Health Records

Sayantee Jana, IIT Hyderabad

Medicine has a big data problem. National healthcare registries provide invaluable data about patients and the reasons why they visit clinics and hospitals. Unfortunately, there's so much of it, and relatively little has been done to organize it in a way that's meaningful for understanding patient experiences. That makes probing by traditional research methods virtually impossible. How are researchers addressing this problem? One team has drawn inspiration from genomics. Using an algorithm routinely applied to study co-expression patterns of multiple genes, they've identified patterns in the circumstances surrounding people experiencing a traumatic brain injury. This approach could help clinicians spot markers of an impending injury before it happens.

The team started with population-wide healthcare data spanning 9 year for residents of Ontario, Canada. According to that information, nearly a quarter million patients visited an emergency department or acute care setting for a first-time traumatic brain injury (or TBI) during that period. Studies have recently established that TBI can be a result of various pre-existing conditions that can modify the risk of injury. These include depressive and substance-use disorders, vascular disease, and medication effects - each of which can itself be affected by factors such as age and socioeconomic status. That makes for a vastly complex web of possible associations that is humanly impossible to untangle. Computationally, however, the task is feasible.

For this, the team used an approach called multiple testing. As its name suggests, the technique simultaneously tests thousands of connections between bits of data, and is normally used to identify significant associations in genetic research. Multiple testing grouped a total of 2600 codes used by healthcare providers to classify patients' conditions into categories. Then, using factor analysis, those categories were whittled down to 43 that were significantly related to TBI versus non-TBI hospital visits. Sorted by effect size, the factors topping the list included those linked to environmental exposures, assault and child abuse, and the adverse effects of medications and drugs in the years leading up to TBI. These findings suggest that such factors could be critical in assessing patients' susceptibility to brain injury. And they point to the complexity of social circumstances surrounding an individual. The team's method provides a way of mining the piles of useful healthcare data to prevent injury and deliver precision medicine. These methods can be easily replicated for medical conditions other than TBI or in sectors other than healthcare, wherever, big data is available, including finance, sales etc.