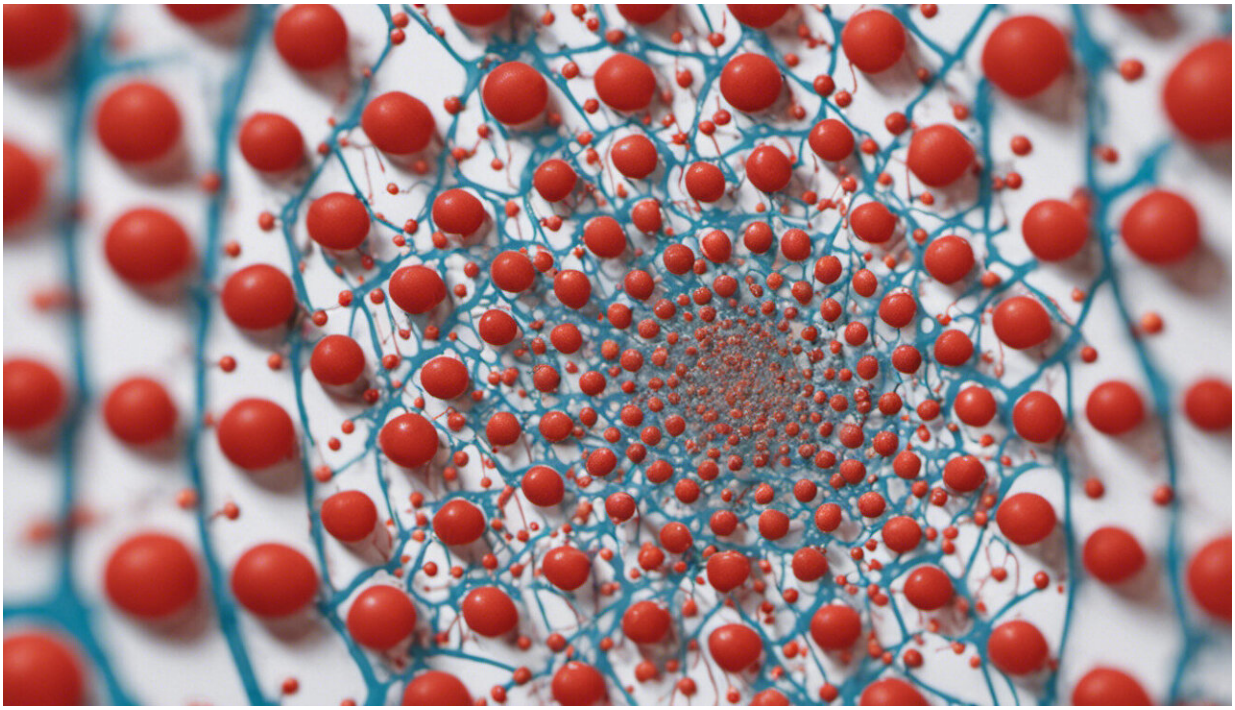
Harvard undergrad's AI model helps to predict TB resistance

May 3 2019, by Ekaterina Pesheva



Credit: AI-generated image (disclaimer)

One of the greatest challenges in treating tuberculosis—the top infectious killer worldwide, according to the World Health Organization (WHO)—is the bacterium's ability to shapeshift rapidly and become resistant to multiple drugs. Identifying resistant strains quickly and choosing the right antibiotics to treat them remains difficult for several

reasons, including the bacterium's propensity to grow slowly in the lab, which can delay drug-sensitivity test results by as much as six weeks after initial diagnosis.

New tests that can quickly and reliably detect resistance to the most commonly used drugs before a patient begins treatment are urgently needed to improve outcomes and help curb the spread of the infection.

Now, Harvard College applied math student Michael Chen '20, working with biomedical researchers at Harvard Medical School's Blavatnik Institute, has designed a computer program that sets the stage for the development of such tests.

The program, described April 29 in EBioMedicine, can accurately predict a TB strain's resistance to 10 first- and second-line drugs in a tenth of a second and with greater precision than similar models.

If incorporated into clinical tests, the model could make resistance detection both faster and more accurate, overcoming deficiencies in current resistance-testing methods that either take too long to yield definitive results or are unreliable.

"Drug-resistant forms of TB are hard to detect, hard to treat, and portend poor outcomes for patients," said senior study author Maha Farhat, assistant professor of biomedical informatics at Harvard Medical School and a pulmonary medicine specialist at Massachusetts General Hospital. "The ability to rapidly detect the full profile of resistance upon diagnosis is critical both to improving individual patient outcomes and in reducing the spread of the infection to others."

The new model will be available online soon as an added feature to Harvard Medical School's genTB tool, which analyzes TB data and predicts TB drug resistance.

More than 10 million new cases of TB are diagnosed each year worldwide, according to WHO. Of these new infections, 4 percent are resistant to at least two drugs—a form of the disease known as multidrug-resistant TB, or MDR-TB. Of the drug-resistant infections, one in 10 are extensively drug-resistant, or XDR-TB, and show great resistance to multiple medications. First-line drugs are given as soon as infection is established, while second-line treatments are added if the patient shows symptoms of resistance or a test shows that the bacterial strain is impervious to first-line treatments.

Imperfect testing

In the developing world, where drug-sensitivity testing can be difficult to obtain, many people diagnosed with TB are treated empirically, based on little more than educated clinical guesswork and assumption.

Fewer than half of the countries with a high prevalence of MDR-TB infections have modern diagnostic capabilities. Even in the best-equipped laboratories, conventional, culture-based drug sensitivity testing takes weeks or months before results can be reported because the TB bacterium tends to grow slowly in dishes, a downside magnified by the small but real infection risk faced by lab workers who handle TB samples.

New molecular tests that scan the DNA of TB samples for resistance genes are increasingly being used, but they have their own limitations—the most important being that they detect resistance for only up to four drugs, and not reliably. They also cannot detect the presence of rare genetic variants that give rise to resistance.

The best way to detect all resistance-conferring mutations is to perform an analysis of a bacterium's full genome. There are several drug-sensitivity tests based on whole-genome sequencing data, and while they

detect resistance to all first-line drugs, they tend to perform poorly on resistance detection to second-line drugs, the researchers said.

These tests also tend to classify a mutation in a simple, binary fashion—either as conferring drug resistance or not—which makes them unreliable detectors of drug resistance in strains that contain rare mutations or mutations of unknown significance. Another critical blind spot of current models using whole-genome data is their inability to assess the interactions between various genes and genetic mutations; in other words, how the presence of one gene or a gene variant influences the function of another, a phenomenon known as epistasis. The new model overcomes this flaw.

Beyond simple detection

To solve the challenge, the scientists set out to create a program that could capture the synergistic effects of multiple mutations. They designed and tested five computational models, two of which stood out for their accuracy. One was a statistical model based on a form of logistic regression analysis, a way to assess the effect of one variable on another. The other was a neural network approach that combined two modes of analysis—wide learning and deep learning. The wide-learning part of the model is similar to a statistical model where each mutation is coded as a variable that either confers resistance or doesn't. The deep-learning part incorporates hidden layers that assess the interactive effects of multiple genes and multiple mutations. The result is a model that behaves more or less as a diagnostic tool than can assess all available information against prior knowledge to make a determination about resistance.

"Our goal was to develop a neural network model, which is a type of machine learning that loosely resembles how connections between neurons are formed in the brain," said Chen, who started writing the

program as a first-year and is now a junior pursuing a degree in applied math with a secondary degree in computer science. The study's first author, he plans to go to medical school.

"The wide and deep neural network interlaces two forms of machine learning to identify the combined effects of genetic variants on antibiotic resistance," Chen said.

The logistic regression model and the wide-and-deep learning model performed comparably, but the wide-and-deep model had a slight advantage in predicting resistance to several second-line therapies due to its ability to analyze the combined effects of multiple mutations on a strain's resistance, as well as the effects of extremely rare genetic mutations with little or unknown significance. This latter feature renders the wide-and-deep learning model more accurate in its predictions than previous models, the research team said, giving it the ability to capture resistance even in samples that contain rare gene mutations whose influence on drug sensitivity is not well understood. In contrast, previously developed machine-learning platforms rely on spotting common gene mutations that are already known to cause resistance.

"The wide-and-deep model is a decision-making tool that combines all of the data with prior biological knowledge that resistance is caused both by large individual mutations and the interactions between many different mutations," said study co-senior author Andy Beam, faculty member in biomedical informatics at HMS and a visiting lecturer at the Harvard T.H. Chan School of Public Health.

Quality education

What made the models successful? Researchers credit the richness of the data on which the programs were "trained." Rather than learning a predefined set of common genetic mutations that are known to promote

resistance, the models were exposed to a data set that included a wide range of genetic mutations with a variety of gene insertions and deletions and extremely rare variants appearing only in a few of the isolates or even one isolate.

The models were trained on 3,601 TB strains resistant to first- and second-line drugs, including 1,228 multidrug-resistant strains. The data set included results from drug-sensitivity testing, allowing the model to assess the link between the presence of certain mutations and drug sensitivity. The training sets were generated with data from various countries curated by researchers at the Critical Path Institute, with support from the Bill & Melinda Gates Foundation.

To test their performance in a realistic setting, the models were challenged to predict resistance in a set of 792 fully sequenced TB genomes they had never "seen" before. This created a high-fidelity testing scenario that eliminated the chance that the models would encounter familiar strains they had "studied" during training.

The wide-and-deep neural network model predicted resistance to first-line drugs with 94 percent accuracy, on average, and 90 percent accuracy to second-line drugs, on average. The statistical model predicted resistance with 94 percent and 88 percent accuracy, respectively. The wide-and-deep neural network model showed slightly greater accuracy in predicting resistance to three specific drugs.

Both models are capable of predicting resistance to first- and second-line therapies in a tenth of a second, underscoring the increasingly important role that artificial intelligence and big data will play in the rapid detection and choice of therapy for the disease.

"Our model highlights the role of artificial intelligence in the case of TB, but its importance goes well beyond TB," Beam said. "AI will help guide

clinical decision-making by rapidly synthesizing large amounts of data to help clinicians make the most informed decision in many scenarios and for many other diseases."

This story is published courtesy of the Harvard Gazette*, Harvard University's official newspaper. For additional university news, visit* Harvard.edu.

Provided by Harvard University