

Artificial Intelligence can spot when correlation really does mean causation

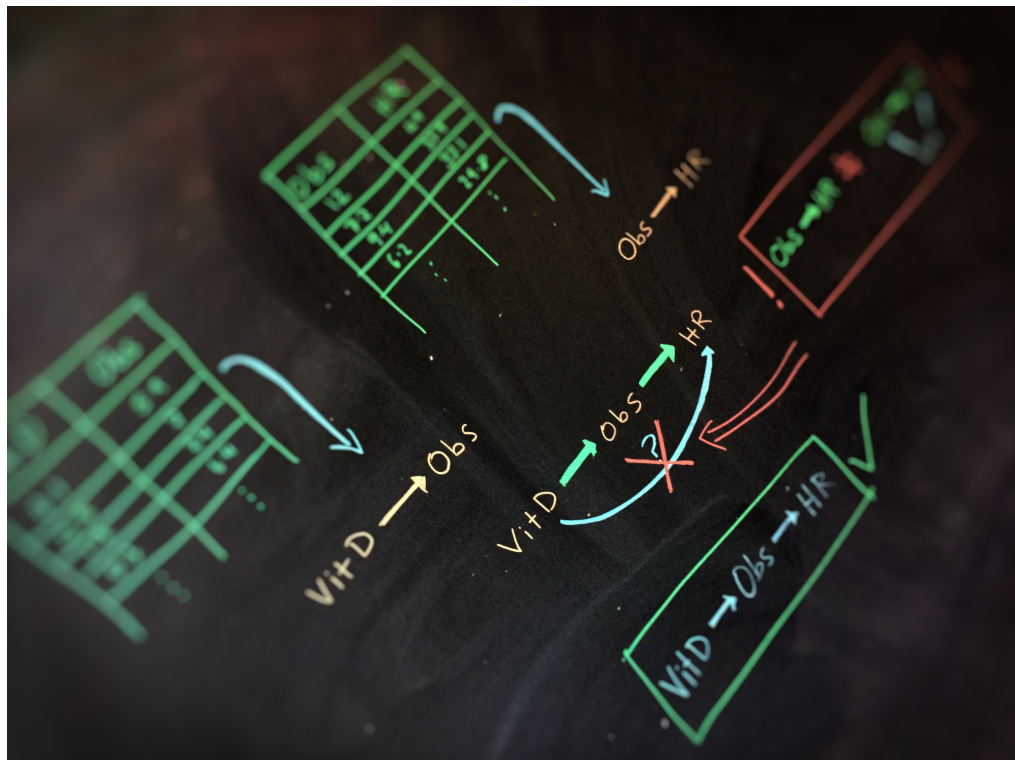
6/2/2020

Peer-reviewed

Modelling & Observational Study

Medical data

- AI can merge overlapping and incomplete medical datasets and then determine which variables are causative, giving new possibilities for old data
- Scientists at Babylon demonstrated the potential of the AI on data from tumors and protein structures



A new Artificial Intelligence (AI) has allowed AI researchers, for the first time, to demonstrate a useful and reliable way of sifting through masses of correlating data to spot when correlation means causation. By fusing old, overlapping and incomplete datasets this new method, inspired by quantum cryptography, paves the way for researchers to glean the results of medical trials that would otherwise be too expensive, difficult or unethical to run. The research is being published at the prestigious and peer-reviewed Association for Advancement of Artificial Intelligence (AAAI) conference in New York.



babylon

Dr Saurabh Johri, Chief Science Officer at Babylon, said: “Until now, we have been limited to piecing together answers from studies that needed to capture all the data really neatly. But when we’ve seen a correlation between obesity and low vitamin D in one study, and obesity and heart failure in another, we have not been able to say whether vitamin D has a causal role in heart failure without doing another, hugely expensive clinical trial. Now we can put the pieces of the jigsaw together.”

Dr Ciarán Lee, Senior Research Scientist at Babylon and Honorary Senior Research Associate at UCL, explained “Scientists have it hammered into them that correlation does not mean causation; ice-cream sales don’t cause sunburn despite rates of both shooting up during the summer. To find the exact cause of sunburn we whittle down or control as many variables as possible. Then when our datasets show that a change in sun exposure matches a change in sunburn, we can be confident the sun exposure was the causative variable. The problem is the real world is rarely neat and tidy and it can be really hard to control all the variables and work out which is causative.”

Scientists started looking for other ways to help spot causative variables. A theory born from physics suggests that everything becomes more disordered and complicated with time, so a cause should be less disordered and complex than its effect. Dr Lee said “If you take your dataset and give each of the variables a complexity rating you can work backwards and spot which one is the cause. But that just helps for that one dataset - we wanted to see if there was a way of combining datasets, ones with gaps or where researchers were asking different questions to what they’re interested in now. That could be a game-changer.”

Dr Lee was inspired by quantum cryptography. The strange laws of quantum physics mean that two users can send a message and then use a mathematical formula to prove whether someone else is eavesdropping on their conversation. Dr Lee realised that datasets could work in a similar way, but thinking of a potential causative variable from another dataset as the eavesdropper. “If one dataset shows us that obesity causes heart disease, and another shows vitamin D causes obesity we can use a mathematical formula to prove whether vitamin D causes obesity or not. This is what our AI is doing.”

“We combined multiple correlating variables from incomplete medical datasets and showed, with a high degree of confidence, which correlations mean causation” said Dr Lee. “I am genuinely excited at what this AI can do. This obviously isn’t a magic wand that will give us all the answers but there are so many studies with missing data, where researchers wish they had tested for something else and could combine it with a study someone else had done, or had thought to ask their questions in a different way. Now they can. Whether it’s the effectiveness of cancer drugs, impact of statins or antidepressants, pesticides or air pollution, the AI should be able to cope with it all.”

The researchers tested the AI on breast cancer and protein-signalling datasets, along with synthetic datasets that were designed to be particularly complex. In each case, the AI found the



babylon

causative variable. In one case it assessed two separate breast tumour datasets, one measuring the perimeter of a breast tumour and the other its texture, and correctly reported that neither caused the other - instead they were both caused by whether the tumour was malignant or benign. Similarly, the AI also determined the signalling structure between two collections of proteins, even whilst missing joint data from a number of the proteins in each dataset.

The algorithm used in the research is available in the paper and on the open access site arXiv so that scientists across the world can use it to reassess overlapping and incomplete datasets. The datasets that were tested are all open-access so that other scientists can verify the research.

“Integrating overlapping datasets using bivariate causal discovery” by Anish Dhir and Ciarán M. Lee will be published at the Association for Advancement of Artificial Intelligence on 7th February 2020

An earlier, non peer-reviewed draft is available on arXiv here:
<https://arxiv.org/pdf/1910.11356.pdf>

For interviews with the authors, or for further information, please contact:

Dr Edward Sykes
(+44) 7966 081 090
ed.sykes@babylonhealth.com
PR & Comms Lead
Babylon

About Babylon

Babylon is a healthtech company with the ambitious mission to put an accessible and affordable health service in the hands of every person on Earth.

Employing hundreds of scientists, doctors and engineers, Babylon combines technology and medical expertise to bring doctors and people closer together. Through a range of digital services - such as AI-backed digital health tools and video doctor appointments - we give people round-the-clock access to affordable (often free) holistic healthcare services. We work with governments, health providers and insurers across the globe, and support healthcare facilities from small local practices to large hospitals.

We have over 3.6 million users and have completed over 1.2 million consultations globally. Some of our partners include the UK's National Health Service, Samsung, Bill & Melinda Gates Foundation and Bupa.