COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Jayavanth Shenoy (Onai) January 30, 2024

Title: Scaled Medical Records Analysis

Guha Jayachandran CIC Database Profile

NSF Award #: 2028008

YouTube Recording with Slides

Winter 2024 CIC Webinar Information

Transcript Editor: Lauren Close

Transcript

Slide 1

My name is Jayavanth and I'm an engineer at Onai. Today, I'll be talking about our project on scaled analysis of medical records with funding from NSF.

Slide 2

The problem is that medical records are very important information, but we don't have a good way of studying them at scale. Traditionally, studies happened in individual hospitals on records within that hospital. Even there, it's laborious, with residents and students manually transcribing data into spreadsheets. This process is also really slow. In situations like COVID-19, this is unacceptable - we don't want to wait until one hospital has enough data for statistically significant studies to happen. That's what wound up happening - I mean, that's what winds up happening usually. But we really want to get results at early stages of the spread. For that, we need to be able to do studies across a number of hospitals in aggregate. There are really basic questions that you think would be easy to answer in this day and age, but they were not so easy for COVID and still aren't even today. For example, does blood pressure medication improve or worsen COVID? Tens of millions of Americans are taking these blood pressure medications, so you would think that this would be an easy question to answer just based on outcomes for patients who take blood pressure medication and wound up in the hospital with COVID. It should be as simple as a Google search query, but it's not.

Slide 3

So our goal was to enable rapid search over medical records across a number of healthcare institutions in an efficient manner and without sacrificing privacy. With our tools, healthcare institutions and pharmaceutical companies can reach statistically significant populations much quicker and research can be conducted right away. It's also more secure. So we can get the best of both worlds? Can we get a huge pool of data for research and at the same time, maximize security?

Slide 4

Yes, we can with this cryptographic technology. The technology to achieve this goal is a magical technique called multi-party computation. This allows a set of parties to jointly computer result over their inputs without revealing their inputs to each other or anyone else. One example of this is called Yao's millionaire problem. Andrew Yao introduced this in 1982 and it's our two millionaires Alice and Bob. They want to know which one of them is richer without revealing their actual wealth. So you have two people, Alice and Bob. They know their own wealth but don't want to tell the other person. And let's say they don't want to involve a third party, so no IRS or anything. Can you think of a way for them to figure out which of them is richer? Well, believe it or not, it turns out there is a way to do this. As another example, let's say we want to calculate the average social security numbers of everyone in this webinar but we don't want to reveal our social security number to each other. Is there a way we can do it? Yes, there is with the secure MPC. We applied the same concept to health records and the result of a query is computed across multiple hospitals without the hospitals needing to transmit or reveal any records. We demonstrated this across a few disparate sites in our project.

Slide 5

Here are some of the examples of queries one might be interested in asking, like the average length of stay in the hospital for patients over 70, or what percentage of very young patients wound up on a ventilator.

Slide 6

For a more cryptographically savvy audience, we used a mascot for NPC technique. I won't go into details now, but you can ask me about this later if you're interested. The only thing I'll mention here is that this protocol is secure against the dishonest majority, which means that this protocol is secure even if a majority of the participants are malicious.

Slide 7

I want to reiterate that with our approach, no data ever leaves the hospital. Each hospital's data is completely siloed in its department. IT departments don't have to worry about security, not just that nothing can be reverse engineered. Coordination of the computations is via blockchain or distributor ledger.

Slide 8

This framework also enables more equitable clinical trials. Traditionally, patient populations and clinical trials have not been very diverse. One of the priorities of the FDA and others is to improve that and with this technology, sites with patients of interest can be located, even if they're not institutions that a company typically works with.

Slide 9

Since then, we have extended this approach from querying to training AI models across disparate sites in a privacy-preserving manner. We are also working with the NIH for privacy-preserving drug discovery. Our end vision here is to have privacy-preserving intelligence networks across biomedicine.

Slide 10

Thank you for listening and I would like to thank CIC for this opportunity and NSF for funding our project, our collaborators across the country and my colleagues at Onai. Please let me know if you have any questions at the end of the session. Thank you again.