COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Murat Kantarcioglu (University of Dallas at Texas), April 15, 2022

Title: Collaborative: A Privacy Risk Assessment Framework for Person-Level Data Sharing During Pandemics

Kelly Dunning CIC Database Profile

NSF Award #: 2029661

YouTube Recording with Slides

April 2022 CIC Webinar Information

Transcript Editor: Saanya Subasinghe

Transcript

Murat Kantarcioglu:

Slide 1

Thank you so much for inviting me back. So when we gave the first talk we had just started this project. Now we are going to conclude it soon. So I'm happy to be able to have, I guess, the record of what our vision is, and what we have done. So I'm going to talk about a specific work that came out of this RAPID project on how to share public data, public health data, while preserving privacy during the pandemic.

Slide 2

So it's clear that sharing surveillance for a data driven response is very very important. We use data for understanding how transmission happens. The data is used for estimating different interventions, what their impact would be, and also, of course, in many cases, and for future pandemics, we will be needing it to detect outbreaks early on.

Slide 3

Now the question is whether we can directly share this data, especially whether we can directly share individual patient level data, which would be useful for building different models. During the Coronavirus crisis, we had also had some kind of data crisis in a sense. Organizations were reluctant to share it and they were concerned about privacy, rightfully so. And this required them to carefully, and in a time-consuming way, analyze which data is to be shared, in what format, and how it can be made public for future use.

Slide 4

So, and this is and one of the compounding challenges, is that unlike traditional data sharing settings we have a data set size that's changing every day. Because every day we may have new patients who can be diagnosed and these new patients' data may need to be shared. And also for regulation purposes, there are additional challenges if you do HIPAA compliance, which is the privacy legislation that governs healthcare data. So some people are very comfortable with HIPAA Safe Harbor rules, which guarantees some certain regular way to sanitize data but the dates are not allowed, and then this creates challenges for some of these end users. And of course, due to the emergency legislation we have to do this very fast. We have to share the data fast without really relating to privacy.

Slide 5

So, in a sense, we developed a framework where we can adapt to these number of records, number of patient records that are changing every day. And we can prioritize different specific information. Let's say you want to be more detailed with respect to age, but not less race but maybe want to be more detailed on race, etc., while understanding privacy implications.

Slide 6

So we developed this risk estimation - privacy risk estimation framework. And the first part of it is that we looked into data generalization. In this work, we focused on the tools where we share the actual correct data that's being given, but at a less specified level or more generalized level.

Slide 7

So what does marginalization mean? It is that, for example, in our framework instead of for privacy reasons, instead of sharing someone's age you may share the age range. Like it says 'five to ten' or you can, if you want to protect privacy even more, you can share a higher range and it may go up to the top where you don't share any information. Of course the leaf [?] nodes are very precise, but the more privacy potential privacy issues, so less privacy protection. And when we go higher up, less information but more privacy protection.

Slide 8

So the second thing is that in order to estimate the risks we really look into the population distribution in different counties and whether, especially for the risk we estimate in this work, called re-identification risk. In other words, an attacker who knows some information about the patients - can they re-identify the data and know that: 'oh, this record must belong to Murat' or 'the second must belong to John.' So in order to do this estimation we look at the census data and use the population distribution to identify it. The next setting is that once we get this data, the time series cases, like how many cases reported, the privacy risk metric we will be using a specific one which I will describe in a second. And also how often, which we call 'windows science,' how often we would like to share the patient records. We created this Monte Carlo Simulation framework where we randomly select the population, we estimate the risk, and we do this thousands of times to estimate this look - at the risks. And here we look at something called PK11 risk, and we want to be less than one percent, which means that the percentage of records falling into the demographic group of size 10 or smaller should be less than or equal to one percent. In other words, we are estimating that less than one percent of the population would be in a group of patients

less than size - total size 11 or less than other 10 on other records. So given this risk estimation, and this is kind of based on what CDC is using, so we try to look into that risk basically used by CDC. And we look at the distribution, and based on these distributions we relate the privacy registrations and the policies.

Slide 9

So in the experiments next I'm going to show, we use this PK 11 list, as I mentioned. We run the simulations 1,000 times and we look into 96 alternative policies. And we do this across the counties and we do this for each county by size and the number of cases.

Slide 10

So what we get is that for small counties when the epidemic starts and we have few cases, privacy risks are much higher than the accepted threshold we mentioned. So you can't really share any data. But as the time progresses, even in small counties you won't be able to share much, but in bigger ones, at least from this risk point of view, you could have many policies. For example this diagram says that if the count is between 1,000 to 50,000 [people] range and we hit the total 5,000 cases, we would be able to find among the 96 policies we looked at we would get 31 to satisfy the risk. And these policies are listed, some of them here, like how fine grained the shared age, whether we had sex, nationality, race, and so on.

Slide 11

In addition to that, we look into dynamic policy change. In other words, we don't stick to changing - to sharing one type of data but we evolve what we share all the time and we compare this with also CDC static policies. In the CDC's case, it divides the age into 0-9 [years old], 10-19, and so on. This kind of intervals, like 10 year intervals. It has combined range and ethnicity, gender, state of residence, and county of residence, and date of first specimen collection. So that's the CDC static policy in terms of data sensitizations used. Here, we looked at whether our dynamic policy, which adapted based on the risk, could perform better. Especially for, we do like, we do daily and weekly releases, basically.

Slide 12

So I won't go into all the details but what happens is that static policies in most of the same cases, whether it's a small county or a big county, turn out to have more number of releases, where the risk privacy threshold is exceeded. So, for example, when we look at the 95 percent quantile for a small county with population size less than 1,000 we would be having for the period we look at, we would have 22 days that the risk is above the threshold. This is daily releases. But for dynamic policy we had even zero. And of course for one million, again, you see the same threshold. So, this kind of showed that one policy about data released and what to format it is may not be good and we need to really adjust as the pandemic evolves.

Slide 13

So in this study what we try to show you is that our dynamic privacy risk assessment framework can give much better results in terms of estimating privacy risks. And it can really adapt to changing environments which protect with better privacy and utility options. But, of course, this work that now we are continuing only looks at the privacy risk. We didn't look into what's the different utility of these policies. In other words, in some scenarios where given privacy is acceptable privacy risk, we have 40 different

policies. But given the new tasks, which policy is better, for example, for outbreak detection, or which policy is better for understanding whether the outbreak is happening in some race, for example. So we didn't really look into those very carefully.

Slide 14

So I will stop here. Again, I would like to thank NSF for supporting us. And this is a joint work with Vanderbilt Medical School and also a colleague from IBM. And this is what I presented in a very short amount of time. If you want more details, it's published in the Journal of the American Medical Informatics Association just recently. So I'll stop here and then towards the end, any questions I will answer live online thank you.