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: Shikhar Johri

Transcript

मूरत कांटार्सिओग्लू:

स्लाइड *1*

मुझे वापस आमंत्रित करने के लिए बहुत-बहुत धन्यवाद। इसलिए जब हमने पहली बात दी तो हमने इस परियोजना को शुरू ही किया था। अब हम इसे जल्द ही समाप्त करने जा रहे हैं। इसलिए मुझे खुशी है कि मैं खुश हूं, मुझे लगता है, हमारी दृष्टि क्या है, और हमने क्या किया है, इसका रिकॉर्ड है। इसलिए मैं एक विशिष्ट कार्य के बारे में बात करने जा रहा हूं जो इस रैपिड परियोजना से निकला है कि महामारी के दौरान गोपनीयता को संरक्षित करते हुए सार्वजनिक डेटा, सार्वजनिक स्वास्थ्य डेटा कैसे साझा किया जाए।

स्लाइड *2*

इसलिए यह स्पष्ट है कि डेटा संचालित प्रतिक्रिया के लिए निगरानी साझा करना बहुत महत्वपूर्ण है। हम डेटा का उपयोग यह समझने के लिए करते हैं कि ट्रांसमिशन कैसे होता है। डेटा का उपयोग विभिन्न हस्तक्षेपों का अनुमान लगाने के लिए किया जाता है, उनका प्रभाव क्या होगा, और निश्चित रूप से, कई मामलों में, और भविष्य की महामारियों के लिए, हमें प्रकोपों का जल्दी पता लगाने के लिए इसकी आवश्यकता होगी।

स्लाइड *3*

अब सवाल यह है कि क्या हम इस डेटा को सीधे साझा कर सकते हैं, खासकर क्या हम सीधे व्यक्तिगत रोगी स्तर के डेटा को साझा कर सकते हैं, जो विभिन्न मॉडलों के निर्माण के लिए उपयोगी होगा। कोरोना वायरस संकट के दौरान एक मायने में हमारे सामने किसी तरह का डेटा संकट भी आया था। संगठन इसे साझा करने के लिए अनिच्छुक थे और वे गोपनीयता के बारे में चिंतित थे, ठीक है। और इसके लिए उन्हें

सावधानीपूर्वक और समय लेने वाले तरीके से विश्लेषण करने की आवश्यकता थी कि कौन सा डेटा साझा किया जाना है, किस प्रारूप में, और भविष्य में उपयोग के लिए इसे कैसे सार्वजनिक किया जा सकता है।

स्लाइड *4*

तो, और यह है और जटिल चुनौतियों में से एक है, यह है कि पारंपरिक डेटा साझाकरण सेटिंग्स के विपरीत हमारे पास एक डेटा सेट आकार है जो हर दिन बदल रहा है। क्योंकि हर दिन हमारे पास नए रोगी हो सकते हैं जिनका निदान किया जा सकता है और इन नए रोगियों के डेटा को साझा करने की आवश्यकता हो सकती है। और विनियमन उद्देश्यों के लिए भी, यदि आप HIPAA अनुपालन करते हैं तो अतिरिक्त चुनौतियाँ हैं, जो कि गोपनीयता कानून है जो स्वास्थ्य देखभाल डेटा को नियंत्रित करता है। इसलिए कुछ लोग एचआईपीएए सेफ हार्बर नियमों के साथ बहुत सहज हैं, जो डेटा को साफ करने के कुछ निश्चित नियमित तरीके की गारंटी देता है लेकिन तारीखों की अनुमति नहीं है, और फिर यह इन अंतिम उपयोगकर्ताओं में से कुछ के लिए चुनौतियां पैदा करता है। और निश्चित रूप से, आपातकालीन कानून के कारण हमें यह बहुत तेजी से करना होगा। हमें वास्तव में गोपनीयता से संबंधित बिना डेटा को तेजी से साझा करना होगा।

स्लाइड *5*

इसलिए, एक अर्थ में, हमने एक ढांचा विकसित किया जहां हम इन रिकॉर्डों की संख्या, रोगी रिकॉर्ड की संख्या को अनुकूलित कर सकते हैं जो हर दिन बदल रहे हैं। और हम विभिन्न विशिष्ट जानकारी को प्राथमिकता दे सकते हैं। मान लीजिए कि आप उम्र के संबंध में अधिक विस्तृत होना चाहते हैं, लेकिन कम दौड़ नहीं बल्कि गोपनीयता निहितार्थ को समझते हुए शायद दौड़ आदि पर अधिक विस्तृत होना चाहते हैं।

स्लाइड *6*

इसलिए हमने इस जोखिम अनुमान - गोपनीयता जोखिम अनुमान ढांचे को विकसित किया है। और इसका पहला भाग यह है कि हमने डेटा सामान्यीकरण पर ध्यान दिया। इस काम में, हमने उन उपकरणों पर ध्यान केंद्रित किया जहां हम वास्तविक सही डेटा साझा करते हैं जो दिया जा रहा है, लेकिन कम निर्दिष्ट स्तर या अधिक सामान्यीकृत स्तर पर।

स्लाइड *7*

तो हाशिए का क्या मतलब है? यह है कि, उदाहरण के लिए, गोपनीयता कारणों के बजाय हमारे ढांचे में, किसी की उम्र साझा करने के बजाय आप आयु सीमा साझा कर सकते हैं। जैसे यह 'पांच से दस' कहता है या आप कर सकते हैं, यदि आप गोपनीयता की और भी अधिक रक्षा करना चाहते हैं, तो आप एक उच्च श्रेणी साझा कर सकते हैं और यह शीर्ष पर जा सकता है जहां आप कोई जानकारी साझा नहीं करते हैं। बेशक पत्ती [?] नोड्स बहुत सटीक हैं, लेकिन अधिक गोपनीयता संभावित गोपनीयता मुद्दे, इसलिए कम गोपनीयता सुरक्षा। और जब हम ऊपर जाते हैं, तो कम जानकारी लेकिन अधिक गोपनीयता सुरक्षा।

स्लाइड *8*

तो दूसरी बात यह है कि जोखिमों का अनुमान लगाने के लिए हम वास्तव में विभिन्न काउंटियों में जनसंख्या वितरण को देखते हैं और क्या, विशेष रूप से जोखिम के लिए हम इस काम में अनुमान लगाते हैं, जिसे पुन: पहचान जोखिम कहा जाता है। दूसरे शब्दों में, एक हमलावर जो रोगियों के बारे में कुछ जानकारी जानता है - क्या वे डेटा को फिर से पहचान सकते हैं और जान सकते हैं कि: 'ओह, यह रिकॉर्ड मूरत का होना चाहिए' या 'दूसरा जॉन का होना चाहिए। इसलिए इस अनुमान को करने के लिए हम

जनगणना के आंकड़ों को देखते हैं और इसकी पहचान करने के लिए जनसंख्या वितरण का उपयोग करते हैं। अगली सेटिंग यह है कि एक बार जब हम यह डेटा प्राप्त कर लेते हैं, तो समय श्रृंखला के मामले, जैसे कि कितने मामलों की सूचना दी जाती है, गोपनीयता जोखिम मीट्रिक हम एक विशिष्ट का उपयोग करेंगे जिसका मैं एक सेकंड में वर्णन करूंगा। और यह भी कि कितनी बार, जिसे हम 'विंडोज़ साइंस' कहते हैं, हम कितनी बार रोगी रिकॉर्ड साझा करना चाहते हैं। हमने इस मोंटे कार्लो सिमुलेशन ढांचे को बनाया जहां हम बेतरतीब ढंग से आबादी का चयन करते हैं, हम जोखिम का अनुमान लगाते हैं, और हम इस रूप का अनुमान लगाने के लिए हजारों बार ऐसा करते हैं - जोखिमों पर। और यहां हम पीके 11 जोखिम नामक कुछ को देखते हैं, और हम एक प्रतिशत से कम होना चाहते हैं, जिसका अर्थ है कि आकार 10 या छोटे के जनसांख्यिकीय समूह में आने वाले रिकॉर्ड का प्रतिशत एक प्रतिशत से कम या उसके बराबर होना चाहिए। दूसरे शब्दों में, हम अनुमान लगा रहे हैं कि एक प्रतिशत से भी कम आबादी आकार से कम रोगियों के समूह में होगी - कुल आकार 11 या अन्य रिकॉर्ड पर अन्य 10 से कम। इसलिए इस जोखिम अनुमान को देखते हुए, और यह सीडीसी द्वारा उपयोग किए जा रहे जोखिम पर आधारित है, इसलिए हम मूल रूप से सीडीसी द्वारा उपयोग किए जाने वाले जोखिम पर गौर करने का प्रयास करते हैं। और हम वितरण को देखते हैं, और इन वितरणों के आधार पर हम गोपनीयता पंजीकरण और नीतियों से संबंधित हैं।

स्लाइड *9*

तो आगे के प्रयोगों में मैं दिखाने जा रहा हूं, हम इस पीके 11 सूची का उपयोग करते हैं, जैसा कि मैंने उल्लेख किया है। हम सिमुलेशन 1,000 बार चलाते हैं और हम 96 वैकल्पिक नीतियों को देखते हैं। और हम इसे काउंटियों में करते हैं और हम इसे प्रत्येक काउंटी के लिए आकार और मामलों की संख्या के अनुसार करते हैं।

स्लाइड *10*

तो हमें जो मिलता है वह यह है कि छोटे काउंटियों के लिए जब महामारी शुरू होती है और हमारे पास कुछ मामले होते हैं, तो गोपनीयता जोखिम हमारे द्वारा उल्लिखित स्वीकृत सीमा से बहुत अधिक होते हैं। तो आप वास्तव में कोई डेटा साझा नहीं कर सकते। लेकिन जैसे-जैसे समय आगे बढ़ता है, यहां तक कि छोटे काउंटियों में भी आप बहुत कुछ साझा नहीं कर पाएंगे, लेकिन बड़े लोगों में, कम से कम इस जोखिम के दृष्टिकोण से, आपके पास कई नीतियां हो सकती हैं। उदाहरण के लिए, यह आरेख कहता है कि यदि गिनती 1,000 से 50,000 [लोगों] की सीमा के बीच है और हम कुल 5,000 मामलों को मारते हैं, तो हम उन 96 नीतियों में से एक पा सकेंगे जिन्हें हमने देखा था, हमें जोखिम को पूरा करने के लिए 31 मिलेंगे। और इन नीतियों को सूचीबद्ध किया गया है, उनमें से कुछ यहां हैं, जैसे कि साझा उम्र कितनी अच्छी है, क्या हमारे पास सेक्स, राष्ट्रीयता, नस्ल, और इसी तरह है।

स्लाइड *11*

इसके अलावा, हम गतिशील नीति परिवर्तन पर गौर करते हैं। दूसरे शब्दों में, हम एक प्रकार के डेटा को साझा करने के लिए बदलने से चिपके नहीं रहते हैं, लेकिन हम हर समय जो साझा करते हैं उसे विकसित करते हैं और हम इसकी तुलना सीडीसी स्थिर नीतियों से भी करते हैं। सीडीसी के मामले में, यह उम्र को 0-9 [वर्ष], 10-19, और इसी तरह विभाजित करता है। इस तरह के अंतराल, जैसे 10 साल के अंतराल. इसमें संयुक्त सीमा और जातीयता, लिंग, निवास की स्थिति और निवास की काउंटी, और पहले नमूना संग्रह की तारीख है। तो यह उपयोग किए गए डेटा संवेदीकरण के संदर्भ में सीडीसी स्थिर नीति है। यहां, हमने देखा कि क्या हमारी गतिशील नीति, जो जोखिम के आधार पर अनुकूलित है, बेहतर प्रदर्शन कर सकती है। विशेष रूप से, हम पसंद करते हैं, हम मूल रूप से दैनिक और साप्ताहिक रिलीज करते हैं।

स्लाइड *12*

इसलिए मैं सभी विवरणों में नहीं जाऊंगा, लेकिन क्या होता है कि अधिकांश मामलों में स्थिर नीतियां, चाहे वह एक छोटी काउंटी हो या एक बड़ी काउंटी, अधिक संख्या में रिलीज होती हैं, जहां जोखिम गोपनीयता सीमा पार हो जाती है। इसलिए, उदाहरण के लिए, जब हम 1,000 से कम जनसंख्या वाले एक छोटे से काउंटी के लिए 95 प्रतिशत मात्रा को देखते हैं, तो हम जिस अवधि को देखते हैं, उसके लिए हमारे पास 22 दिन होंगे, हमारे पास 22 दिन होंगे कि जोखिम सीमा से ऊपर है। यह दैनिक रिलीज है। लेकिन गतिशील नीति के लिए हमारे पास शून्य भी था। और निश्चित रूप से एक मिलियन के लिए, फिर से, आप एक ही सीमा देखते हैं। इसलिए, इस तरह से पता चला कि जारी किए गए डेटा के बारे में एक नीति और इसे क्या प्रारूपित करना है, यह अच्छा नहीं हो सकता है और हमें वास्तव में समायोजित करने की आवश्यकता है क्योंकि महामारी विकसित होती है।

स्लाइड *13*

इसलिए इस अध्ययन में हम आपको जो दिखाने की कोशिश करते हैं वह यह है कि गोपनीयता जोखिमों का आकलन करने के मामले में हमारा गतिशील गोपनीयता जोखिम मूल्यांकन ढांचा बेहतर परिणाम दे सकता है। और यह वास्तव में बदलते परिवेश के अनुकूल हो सकता है जो बेहतर गोपनीयता और उपयोगिता विकल्पों के साथ रक्षा करता है। लेकिन, निश्चित रूप से, यह काम जो अब हम जारी रख रहे हैं, केवल गोपनीयता जोखिम को देखता है। हमने इस बात पर ध्यान नहीं दिया कि इन नीतियों की अलग उपयोगिता क्या है। दूसरे शब्दों में, कुछ परिदृश्यों में जहां दी गई गोपनीयता स्वीकार्य गोपनीयता जोखिम है, हमारे पास 40 अलग-अलग नीतियां हैं। लेकिन नए कार्यों को देखते हुए, कौन सी नीति बेहतर है, उदाहरण के लिए, प्रकोप का पता लगाने के लिए, या कौन सी नीति यह समझने के लिए बेहतर है कि क्या प्रकोप कुछ दौड़ में हो रहा है, उदाहरण के लिए। इसलिए हमने वास्तव में उन पर बहुत ध्यान से नहीं देखा।

स्लाइड *14*

इसलिए मैं यहीं रुक जाऊंगा। फिर से, मैं एनएसएफ को हमारा समर्थन करने के लिए धन्यवाद देना चाहता हूं। और यह वेंडरबिल्ट मेडिकल स्कूल के साथ एक संयुक्त कार्य है और आईबीएम के एक सहयोगी भी हैं। और यह वही है जो मैंने बहुत कम समय में प्रस्तुत किया है। यदि आप अधिक विवरण चाहते हैं, तो यह हाल ही में अमेरिकन मेडिकल इंफॉर्मेटिक्स एसोसिएशन के जर्नल में प्रकाशित हुआ है। इसलिए मैं यहां रुकूंगा और फिर अंत में, किसी भी प्रश्न का उत्तर मैं ऑनलाइन लाइव दूंगा धन्यवाद।