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: Julie Meunier

Transcript

Slide 1

Merci beaucoup de m'avoir réinvité. Lorsque nous avons fait notre premier exposé, nous venions de lancer ce projet. Aujourd'hui, nous allons bientôt le conclure. Je suis donc ravie de pouvoir conserver, je pense, la trace de notre vision et de ce que nous avons fait. Je vais donc vous parler d'un travail spécifique issu de ce projet RAPID sur la manière de partager les données publiques, les données de santé publique, tout en préservant la vie privée pendant la pandémie.

Slide 2

Il est clair que le partage de la surveillance pour une réponse fondée sur les données est très important. Nous utilisons les données pour comprendre comment se produit la transmission. Les données sont utilisées pour estimer les différentes interventions, leur impact, et bien sûr, dans de nombreux cas, et pour les pandémies futures, nous en aurons besoin pour détecter les épidémies à un stade précoce.

Slide 3

La question est maintenant de savoir si nous pouvons partager directement ces données, en particulier si nous pouvons partager directement les données individuelles des patients, qui seraient utiles pour construire différents modèles. Lors de la crise du coronavirus, nous avons également connu une sorte de crise des données. Les organisations étaient réticentes à les partager et s'inquiétaient, à juste titre, du respect de la vie privée. Elles ont donc dû analyser soigneusement, et avec beaucoup de temps, quelles données devaient être partagées, dans quel format et comment elles pouvaient être rendues publiques en vue d'une utilisation future.

Slide 4

L'un des défis majeurs est que, contrairement aux cadres traditionnels de partage des données, la taille de l'ensemble des données change tous les jours. En effet, chaque jour, de nouveaux patients peuvent être diagnostiqués et les données de ces nouveaux patients peuvent devoir être partagées. En outre, pour des raisons de réglementation, la conformité à la loi HIPAA, qui régit la confidentialité des données de santé, pose des problèmes supplémentaires. Certaines personnes sont très à l'aise avec les règles de la sphère de sécurité de l'HIPAA, qui garantissent une certaine régularité dans l'assainissement des données, mais les dates ne sont pas autorisées, ce qui pose des problèmes à certains de ces utilisateurs finaux. Et bien sûr, en raison de la législation d'urgence, nous devons agir très rapidement. Nous devons partager les données rapidement sans vraiment tenir compte de la vie privée.

Slide 5

Dans un sens, nous avons donc développé un cadre qui nous permet de nous adapter au nombre de dossiers, au nombre de dossiers de patients qui changent tous les jours. Et nous pouvons donner la priorité à différentes informations spécifiques. Disons que vous voulez plus de détails sur l'âge, mais pas moins sur la race, mais que vous voulez peut-être plus de détails sur la race, etc.

Slide 6

Nous avons donc élaboré ce cadre d'estimation des risques - d'estimation des risques pour la vie privée. Dans un premier temps, nous nous sommes penchés sur la généralisation des données. Dans ce travail, nous nous sommes concentrés sur les outils qui permettent de partager les données exactes qui sont données, mais à un niveau moins précis ou plus généralisé.

Slide 7

Que signifie la marginalisation ? Cela signifie que, par exemple, dans notre cadre, au lieu de communiquer l'âge d'une personne pour des raisons de confidentialité, vous pouvez communiquer la tranche d'âge. Par exemple, on peut dire " cinq à dix " ou, si l'on veut protéger encore plus la vie privée, on peut partager une fourchette plus élevée et aller jusqu'au sommet où l'on ne partage aucune information. Bien sûr, les nœuds feuilles sont très précis, mais plus il y a de problèmes potentiels de confidentialité, moins il y a de protection de la vie privée. Plus on monte dans la hiérarchie, moins il y a d'informations, mais plus la protection de la vie privée est importante.

Slide 8

Deuxièmement, pour estimer les risques, nous examinons vraiment la distribution de la population dans les différents comtés et si, en particulier pour le risque que nous estimons dans ce travail, appelé risque de ré-identification. En d'autres termes, un attaquant qui connaît certaines informations sur les patients peut-il réidentifier les données et savoir que "oh, ce dossier doit appartenir à un patient" : Oh, ce dossier doit appartenir à Murat" ou "le second doit appartenir à John". Pour réaliser cette estimation, nous examinons les données de recensement et utilisons la distribution de la population pour l'identifier. Une fois que nous avons obtenu ces données, la série chronologique des cas, comme le nombre de cas signalés, la mesure du risque pour la vie privée, nous en utiliserons une spécifique que je décrirai dans un instant. Et aussi la fréquence, que nous appelons la "science des fenêtres", la fréquence à laquelle nous aimerions partager les dossiers des patients. Nous avons créé ce cadre de simulation de Monte

Carlo dans lequel nous sélectionnons au hasard la population, nous estimons le risque, et nous faisons cela des milliers de fois pour estimer ce regard sur les risques. Ici, nous examinons ce que l'on appelle le risque PK11, et nous voulons qu'il soit inférieur à un pour cent, ce qui signifie que le pourcentage de dossiers appartenant au groupe démographique de taille 10 ou inférieure doit être inférieur ou égal à un pour cent. En d'autres termes, nous estimons que moins d'un pour cent de la population ferait partie d'un groupe de patients de taille inférieure à 11 ou de taille inférieure à 10 dans d'autres dossiers. Compte tenu de cette estimation du risque, qui est en quelque sorte basée sur les données du CDC, nous essayons d'examiner le risque utilisé par le CDC. Nous examinons la distribution et, sur la base de ces distributions, nous établissons un lien entre les enregistrements et les politiques en matière de protection de la vie privée.

Slide 9

Dans les expériences que je vais maintenant vous présenter, nous utilisons cette liste PK 11, comme je l'ai mentionné. Nous effectuons les simulations 1 000 fois et nous examinons 96 politiques alternatives. Et nous le faisons pour l'ensemble des comtés et pour chaque comté en fonction de sa taille et du nombre de cas.

Slide 10

Nous constatons que dans les petits comtés, lorsque l'épidémie commence et qu'il y a peu de cas, les risques pour la vie privée sont beaucoup plus élevés que le seuil accepté que nous avons mentionné. Il n'est donc pas possible de partager des données. Mais au fur et à mesure que le temps passe, même dans les petits comtés, vous ne pourrez pas partager grand-chose, mais dans les plus grands, du moins du point de vue du risque, vous pourriez avoir de nombreuses politiques. Par exemple, ce diagramme indique que si le nombre de personnes se situe entre 1 000 et 50 000 [personnes] et que nous atteignons un total de 5 000 cas, nous pourrions trouver parmi les 96 polices que nous avons examinées, 31 pour satisfaire le risque. Ces politiques sont énumérées, certaines d'entre elles ici, comme le degré de précision de l'âge partagé, le sexe, la nationalité, la race, etc.

Slide 11

En outre, nous examinons les changements dynamiques de politique. En d'autres termes, nous ne nous contentons pas de changer - de partager un type de données - mais nous faisons évoluer ce que nous partageons en permanence et nous le comparons aux politiques statiques des CDC. Dans le cas du CDC, l'âge est divisé en 0-9 [ans], 10-19, et ainsi de suite. Ce type d'intervalles, comme des intervalles de 10 ans. Il combine l'étendue et l'ethnicité, le sexe, l'état de résidence, le comté de résidence et la date de la première collecte d'échantillons. Voilà donc la politique statique des CDC en termes de sensibilisation aux données. Ici, nous avons cherché à savoir si notre politique dynamique, qui s'adapte en fonction du risque, pouvait être plus performante. En particulier pour les diffusions quotidiennes et hebdomadaires.

Slide 12

Je n'entrerai pas dans les détails, mais ce qui se passe, c'est que les politiques statiques, dans la plupart des cas, qu'il s'agisse d'un petit ou d'un grand comté, donnent lieu à un plus grand nombre de disséminations, lorsque le seuil de confidentialité du risque est dépassé. Ainsi, par exemple, si l'on considère le quantile de 95 % pour un petit comté dont la population est inférieure à 1 000 habitants, il y aurait, pour la période considérée, 22 jours où le risque est supérieur au seuil. Il s'agit de communiqués

quotidiens. Mais pour la politique dynamique, nous avons même zéro. Et bien sûr, pour un million, on retrouve le même seuil. Cela montre donc qu'une politique unique concernant les données publiées et leur format n'est peut-être pas la bonne et qu'il faut vraiment s'adapter à l'évolution de la pandémie.

Slide 13

Dans cette étude, nous essayons de vous montrer que notre cadre d'évaluation dynamique des risques d'atteinte à la vie privée peut donner de bien meilleurs résultats en termes d'estimation des risques d'atteinte à la vie privée. Et il peut vraiment s'adapter à des environnements changeants qui protègent avec de meilleures options de protection de la vie privée et d'utilité. Mais, bien sûr, ce travail que nous poursuivons actuellement ne porte que sur les risques pour la vie privée. Nous n'avons pas examiné les différentes utilités de ces politiques. En d'autres termes, dans certains scénarios où le risque pour la vie privée est acceptable, nous disposons de 40 politiques différentes. Mais compte tenu des nouvelles tâches, quelle politique est la meilleure, par exemple, pour la détection d'une épidémie, ou quelle politique est la meilleure pour comprendre si l'épidémie se produit dans une certaine course, par exemple. Nous ne nous sommes donc pas penchés sur ces questions avec beaucoup d'attention.

Slide 14

Je vais donc m'arrêter ici. Je tiens à nouveau à remercier la NSF pour son soutien. Il s'agit d'un travail conjoint avec la Vanderbilt Medical School et un collègue d'IBM. Voilà ce que j'ai présenté en très peu de temps. Si vous voulez plus de détails, l'article a été publié récemment dans le Journal of the American Medical Informatics Association. Je vais donc m'arrêter ici et vers la fin, je répondrai en direct à toutes les questions, merci.