COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Dominique Duncan (University of Southern California), January 2021

Title: COVID-ARC (COVID-19 Data Archive)

Dominique Duncan CIC Database Profile

NSF Award #: 2027456

Youtube Recording with Slides

January 2022 CIC Webinar Information

Transcript Editor: Julie Meunier

Transcript

Slide 1

Merci à Florence, à Lauren et à tous les membres de l'équipe COVID Info Commons. J'ai fait un premier exposé il y a environ un an et demi et, à la fin de cet exposé, je ferai le point sur ce qui en a résulté et sur une nouvelle collaboration qui s'est mise en place. Mais je voulais parler de notre archive de données COVID-19, COVID-ARC en abrégé. Je suis professeur adjoint à l'Université de Californie du Sud, à la Keck School of Medicine, dans le laboratoire de neuro-imagerie.

Slide 2

Dans notre laboratoire, nous avons donc beaucoup d'expérience dans la création d'archives de données multimodales à grande échelle, principalement pour les données cérébrales. Mais au début de la pandémie, nous avons pensé que nous pourrions utiliser notre expérience et nos ressources avec toutes ces archives de données pour développer une archive de données COVID-19. Nous avons donc reçu une bourse RAPID de la NSF pour développer ces archives de données appelées COVID-ARC. Ce que nous faisons, c'est regrouper différents types de données COVID-19 ainsi que des ressources et nous avons construit une plateforme d'archives centralisées et en réseau qui stockent, conservent, visualisent et diffusent les données multimodales COVID-19. Nous disposons d'un grand nombre d'ensembles de données provenant du monde entier. Nous avons donc travaillé avec les fournisseurs de données pour mettre au point des accords d'utilisation des données adaptés à leurs besoins. Les métadonnées sont également disponibles sur le site web, de sorte que si les utilisateurs souhaitent demander l'accès à ces données, nous facilitons le processus. Mais ce sont les fournisseurs de données qui prennent la décision finale. Une grande partie des données est stockée sur notre site à l'USC, mais certains ensembles de

données sont stockés sur le site où les données ont été collectées et nous disposons des métadonnées pour que les gens puissent les consulter. Une grande partie de notre travail a consisté à harmoniser les métadonnées, afin de faciliter la recherche sur les cohortes regroupées et de permettre aux chercheurs d'effectuer différents types d'analyses sur différents sites, plutôt que de se concentrer sur un seul d'entre eux. Dans quelques diapositives, je parlerai de certains de ces défis et de la raison pour laquelle nous travaillons sur l'harmonisation. Nous avons également intégré des outils de visualisation, de contrôle de la qualité et d'analyse. Encore une fois, pour aider les chercheurs - juste pour accélérer la recherche sur COVID-19. En plus de tout le travail d'archivage et d'harmonisation des données, nous effectuons également différents types d'analyses sur les données dont nous disposons et nous utilisons les principes de rétroaction et la science des données pour étudier divers aspects de COVID-19.

Slide 3

En ce qui concerne les données, nous disposons de différents types de données - une grande partie d'entre elles se concentre sur les images de tomodensitométrie et de radiographie, mais nous avons aussi des données cliniques qui comprennent les symptômes, les constantes, les comorbidités, les données démographiques, les antécédents du patient, la géolocalisation. Nous disposons également - pour l'imagerie - d'ultrasons et d'IRM, ainsi que de données EEG. Nous avons également fourni des masques pulmonaires, des masques d'infection et des annotations de radiologues. Vous pouvez voir ici que nous utilisons le système de transfert de fichiers crypté à grande vitesse d'IBM, conforme à la HIPAA, appelé ASPERA, qui permet aux fournisseurs de données de transférer facilement des données à COVID-ARC et aux utilisateurs de télécharger des données de COVID-ARC sur leurs ordinateurs.

Slide 4

À l'heure actuelle, nous disposons de 28 ensembles de données provenant du monde entier. Comme vous pouvez l'imaginer, il existe des incohérences dans la dénomination des fichiers, dans le formatage des métadonnées, dans les infrastructures de stockage, ainsi que d'autres différences entre ces ensembles de données. Nous avons donc rassemblé toutes ces données dans une archive centralisée et nous avons veillé à ce que le nom et l'organisation des fichiers soient cohérents, que le formatage des métadonnées soit cohérent et qu'il soit facile de télécharger plusieurs ensembles de données à partir d'un seul endroit à l'aide d'ASPERA.

Slide 5

Voici une capture d'écran d'une partie des données dont nous disposons. Je n'ai pas pu tout mettre sur une seule diapositive, mais vous pouvez voir que si vous allez sur covid-arc.loni.usc.edu, vous pouvez trouver les données dont nous disposons. Vous pouvez voir le numéro du site, l'endroit où les données ont été collectées, les modalités, les formats de fichiers, toutes les métadonnées que nous avons en plus, et ensuite le nombre d'images qu'il y a. Certaines d'entre elles sont réparties entre COVID et non COVID. Nous disposons également d'informations sur la disponibilité publique ou non des données.

Slide 6

Maintenant, je voudrais juste mettre en lumière quelques-uns des projets sur lesquels mes étudiants ont travaillé, ils ont été très productifs et ont fait des recherches vraiment passionnantes. J'ai quelques expériences de recherche NSF pour les étudiants de premier cycle et Aksh Garg est l'un d'entre eux. Il a commencé alors qu'il était au lycée et il est maintenant en première année de licence à Stanford. La semaine dernière, son article a été accepté dans Expert Systems with Applications et il a comparé 40 architectures de réseaux neuronaux convolutifs pour distinguer le COVID du non COVID et il a découvert que le meilleur modèle EfficientNet-B5 donnait une sensibilité et une spécificité extrêmement élevées en termes de précision. Le modèle s'appuyait également sur des caractéristiques cliniques pertinentes, telles que les opacités de verre et les consolidations, qui sont souvent observées chez les patients atteints de COVID.

Slide 7

Un autre projet - Alex Bruckhaus est un autre lauréat de la REU, dont les travaux ont été publiés dans le Journal of Immigrant and Minority Health - lui et d'autres étudiants se sont penchés sur la dynamique de la vaccination en Californie. Ils se sont donc penchés sur ce que l'on appelle l'indice de vulnérabilité sociale (Social Vulnerability Index, SVI), qui comporte quatre thèmes, à savoir le statut socio-économique, la composition du ménage et le handicap, le type de logement et le transport, ainsi que le statut de minorité et la langue. Ils ont constaté que la couverture vaccinale la plus faible se trouvait dans les groupes à forte vulnérabilité. Le statut minoritaire et la langue ont donné lieu à la plus grande disparité de couverture entre les comtés à faible vulnérabilité et ceux à forte vulnérabilité. Je pense donc qu'il s'agit d'un travail très important, en particulier au moment où nous essayons de vacciner une plus grande partie de la population.

Slide 8

Un autre article publié par Alex Bruckhaus et d'autres étudiants portait sur les taux d'infection après le confinement à la suite des réouvertures. Ils ont donc étudié 83 comtés des États-Unis où le nombre de cas de COVID-19 était élevé l'année dernière. Ils ont examiné différents types d'entreprises et ont fait la distinction entre une réouverture totale et une réouverture partielle. Ils ont étudié les variations des taux d'infection avant et après ces réouvertures et ont cherché à savoir quelles entreprises avaient le plus d'impact sur l'augmentation des taux d'infection. Les bars et les salles de sport ont joué un rôle important à cet égard.

Slide 9

Yujia Zhang, qui est mon assistante sur ce projet, a réalisé l'année dernière un article de synthèse sur l'association entre le groupe sanguin et le COVID-19. Elle a donc examiné 23 études qui présentaient une vue d'ensemble du groupe sanguin en tant que facteur de risque et de protection, de la manière dont certains groupes sanguins sont susceptibles d'être testés positifs et des résultats cliniques de la gravité. Elle a également examiné les associations génétiques et les mécanismes moléculaires sous-jacents potentiels.

Slide 10

Azrin Khan, qui a bénéficié d'une bourse REU au cours des deux derniers étés, travaille sur un projet de segmentation des poumons basée sur un seuil. Il s'agit d'une méthode de seuillage en plusieurs étapes qui permet de quantifier les anomalies pulmonaires avec de meilleures performances que les méthodes existantes.

Slide 11

Et comme je manque de temps, je vais passer un peu vite, mais je voulais parler d'une collaboration qui a commencé à la suite de la première conférence COVID Info Commons que j'ai donnée. Michael Pazzani et Albert Hsiao de l'UCSD, de San Diego, ont également donné des conférences et nous avons entamé une collaboration après cela. Nous avons également soumis une proposition de santé intelligente qui n'a pas été financée, mais nous l'avons soumise à nouveau en novembre dernier, et nous attendons donc de savoir ce qu'il en est. Mais voici une capture d'écran de l'un des webinaires de sensibilisation que nous avons organisés à l'intention des lycéens du sud de la Californie, et nos étudiants ont collectivement donné des conférences éclair sur leur projet, ce qui a été une grande réussite. Je tenais à vous remercier. Voici le site web du laboratoire, le site web COVID-ARC, envoyez-moi un email[duncand@usc.edu] si vous avez des questions. Je remercie la NSF et le NIH pour leur financement.