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: Saanya Subasinghe

Transcript

Dominique Duncan:

Slide 1

Okay, so thank you Florence and Lauren and everyone behind the COVID Info Commons team. I first gave a talk about a year and a half ago and at the end of this talk I'll give some updates about what resulted from that and a new collaboration that was formed. But I wanted to talk about our COVID-19 Data Archive, COVID-ARC for short. I'm an Assistant Professor at the University of Southern California at the Keck School of Medicine in the Laboratory of Neuroimaging.

Slide 2

So, in our lab we have a lot of experience building large-scale, multi-modal data archives, mainly for brain data. But at the beginning of the pandemic, we thought that we might be able to use our experience and resources with all of those data archives and develop a COVID-19 data archive. So we received - we were awarded an NSF RAPID award to develop this data archive called COVID-ARC. So what we do is we aggregate different types of COVID-19 data as well as resources and we have built a platform of networked and centralized archives that store, curate, visualize and disseminate multimodal COVID-19 data. We have a lot of data sets from around the world. Many of them are publicly available, some are private, and so we've worked with those data providers to come up with data use agreements that are tailored to their needs. And we have the metadata available on the website so if users would like to request access to that we facilitate that process. But then the data providers make that final decision. And a lot of the data are stored at our site at USC but for some of the data sets they're stored at the site

where the data were collected and we have the metadata available so people can see that. And a lot of our work has been around harmonizing the metadata and this is to facilitate research on pooled cohorts to make it easier for people to do different types of analyses across different sites rather than focusing on one site. In a couple slides, I'll talk about some of those challenges and why we work on harmonizing it. And we've also integrated visualization and quality control tools and analytic tools. Again, to help researchers - just to expedite research on COVID-19. And in addition to all the work around the data archiving and harmonizing we are also doing different types of analyses on the data that we have and we're using feedback principles and data science to study various aspects of COVID-19.

Slide 3

So, in terms of the data we have different types of data - a lot of it focuses on just CT images as well as X-Ray but we have clinical data that includes symptoms, vitals, comorbidities, demographics, patient history, geolocation. We also have- for imaging we have ultrasound and MRI as well as some EEG data. And then we've also provided lung masks, infection masks, and radiologist annotations. And here you can see we use IBM's HIPAA compliant encrypted high-speed file transfer system called ASPERA and this is a really easy way for data providers to transfer data to COVID-ARC as well as for users to download data from COVID-ARC to their computers.

Slide 4

And right now we have 28 data sets from around the world. As you can imagine there is some inconsistent file naming across these different data sets, inconsistent metadata formatting, differences in storing infrastructures, as well as other differences across those data sets. So what we've done is we have put all of these together into one centralized data archive and we've made sure that there's consistent file naming and organization, consistent metadata formatting, and ease in downloading several data sets from one location using ASPERA.

Slide 5

And here's just a screenshot of part of the data that we have. I couldn't fit everything in on one slide, but you can just see if you go to covid-arc.loni.usc.edu, you can find what data we have. And this is organized you can see we have the site number, the location where the data were collected, the modalities, the file formats, any metadata that we have in addition to that, and then how many images there are. And some of them are split between COVID and not COVID. And then we also have information on whether or not it's publicly available.

Slide 6

So now, I just wanted to highlight a few of the projects that my students have been working on, they have been very productive and have been doing really exciting research. I have a few NSF research experiences for undergrads and Aksh Garg is one of those. He started while he was in high school and now he's an undergrad - a first-year undergrad at Stanford and just last week his paper was accepted in *Expert Systems with Applications* and here he did a comparison of 40 convolutional neural network architectures to distinguish COVID versus not COVID and he found that the best model EfficientNet-B5

yielded extremely high accuracy sensitivity and specificity. And the model also relied on clinically relevant features such as ground Glass opacities and consolidations, which are often seen in COVID 19 patients.

Slide 7

Another project - Alex Bruckhaus is another REU awardee, so this work was published in the *Journal of Immigrant and Minority Health* and here he and other students were looking at the vaccination dynamics in California. So they looked at something called the Social Vulnerability Index and the SVI has - they looked at four SVI themes including socioeconomic status, household composition and disability, housing type and transportation and minority status and language. And they found that the lowest vaccination coverage was in high vulnerability groups. Minority status and language yielded the largest disparity in coverage between low and high vulnerability counties. So I think that this is really important work, especially as we're trying to get more of the population vaccinated.

Slide 8

Another, another paper that Alex Bruckhaus and other students published was looking at post-lockdown infection rates following re-openings. So they looked at 83 counties across the United States with high COVID-19 case counts last year. And they were looking at different types of businesses and they separated between a full reopening or a partial reopening and they were looking at the infection rate changes before and after those reopenings and seeing which businesses had the biggest effect in a rise in infection rates. And so bars and gyms played an important - a large role in that.

Slide 9

Yujia Zhang, who is my project assistant on this project, she did a review paper last year looking at the blood type association with COVID-19. So she looked at 23 studies that had an overview of blood type as both a risk and protective factors, how - how certain blood types people with certain blood types are susceptible for testing positive and the clinical outcomes of severity. And she also went through the genetic associations and potential underlying molecular mechanisms there.

Slide 10

And then Azrin Khan, who has been an REU fellow the past two summers, she is working on this project looking at threshold-based lung segmentation. So this is the multi-step thresholding method to quantify lung abnormalities with better performance than existing methods.

Slide 11

And since I'm running out of time I'm going to rush through this a little bit but I wanted to talk about a collaboration that started because of that first COVID Info Commons talk that I gave and Michael Pazzani and Albert Hsiao from UCSD, from San Diego, they also gave talks and we started a collaboration after that. We also submitted a smart health proposal which didn't get funded, but we resubmitted this past November so we're waiting about that. But this is just a screenshot of one of the outreach webinars that

we had for high school students in Southern California, and collectively our students gave lightning talks about their project and that was, that was very successful. And I just wanted to thank you. This is the lab website the COVID-ARCwebsite, please email me [duncand@usc.edu] if you have any questions And thank you to both NSF and NIH for the funding.