Leveraging the OHDSI vocabulary to characterize the COVID-19 epidemic using Twitter data and NLP

PRESENTER: **Juan M. Banda**

INTRO:

- It is vital to use any available data resources to try to characterize this pandemic and try to generate insights that are useful for public health officials and scientists. Social media data sources, especially Twitter and Reddit, have recently been proven to be useful resources for the characterization of several disasters and outbreaks in the past. While Twitter data contains many structured data fields, the most interesting and insightful ones are unstructured, making it a non-trivial task to standardize any findings and generate any insights from them. When trying to look at this data from the health informatics context, the need for data standardization is vital. Per design, the ODSHI vocabulary is a collection of external vocabularies and ontologies like SNOMED-CT, MeSh, ICD, RxNorm among hundreds of others. Leveraging this vocabulary alongside social media mining tools, will allow us to extract and standardize7 the unstructured data more effectively, and be able to analyze it more efficiently.

Domain ID	Distinct	Concept Class ID	Distinct	Vocabulary ID	Distinct
Drug	24,045	Clinical Finding	9,144	SNOMED	31,444
Condition	18,373	Brand Name	6,893	MedDRA	8,456
Observation	17,593	Substance	5,231	RxNorm Extension	5,469
Procedure	4,013	LLT	4,790	dm+d	4,598
Geography	2,415	Ingredient	3,870	RxNorm	4,102

Table 1. Number of top five unique concepts captured by domain, concept class, and vocabulary identifiers.

In this work we have shown that leveraging the OHDSI vocabulary for NLP tasks on social media data adds great value when looking into health informatics topics. We also show that people share a considerable amount of health data on Twitter, making it a valuable resource for health research

Take a picture to download the full paper

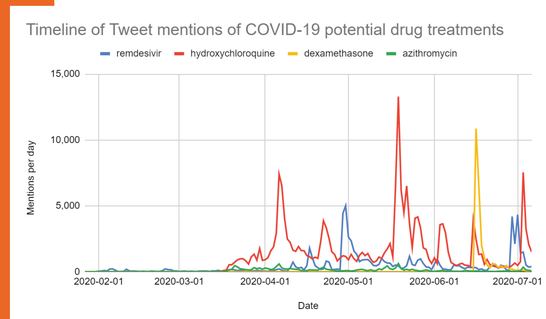
Characterizing drug mentions for COVID-19: a proof of concept

One of the most discussed topics on Twitter has been the discourse around drug treatment. With the produced text annotations, we are able to elucidate time trends when each of the most discussed drugs (Table 2) have gained traction based on news reports, politician mentions, and scientists on Twitter (Figure 1).

Table 2. Drug ingredient mentions found

Drug Ingredient	Frequency
hydroxychloroquine	204,879
remdesivir	72,841
chloroquine	49,915
oxygen	37,961
vitamin D	25,445
dexamethasone	25,142
zinc	24,843
azithromycin	16,079
ibuprofen	8,469
ivermectin	6,390

Notice the individual trends on the drug discourse over time has changed with relevant events, showing first the wide discussion of hydroxychloroquine, touted by many as the first potentially game-changing drug around the beginning of April. This trend reached new heights when the first negative results were published from scientific studies near mid-May. Similar patterns can be seen for remdesivir and dexamethasone later on during the pandemic study period.



Ramya Tekumalla, Juan M. Banda
Georgia State University, Atlanta, Georgia, USA