Q&A from Webinar:

Big Data: Engaging in Public Health Surveillance beyond the Confines of the Local Health Department

Questions for Roderick Jones:

1. A question for Mr. Jones: Why was the decision made to exclude children < 18 yrs. from the data?

I'd like to amend the answer I gave on the webinar. The main reason is that inclusion of pediatric patients was not needed to answer the initial research questions, and their exclusion was thought to reduce the theoretical risk to vulnerable populations with respect to the human subjects research review applications.

2. For the Chicago project, does a data governance committee exist and if so how does it operate relative to public data access?

Yes, we've created a committee consisting of all the site leads with simple majority required for approval of any data access.

3. Given privacy issues: what do you think is the applicability of these techniques when applied to local or even state level analysis with lower populations? What will be available, usable at 250,000 population or less of a city or say counties less than 30,000.

This is an active area of exploration with the other regional extension centers in the state of Illinois (IL-HITREC). We think this will still work, but will require additional roll-up of populations across larger regions with low population density.

4. Have any of the Chicago data led to media backlash or unexpected scrutiny?

Not so far.

Q&A from Webinar:

Big Data: Engaging in Public Health Surveillance beyond the Confines of the Local Health Department

Questions for Mark Dredze:

1. How are you mining the dta from twitter?

We use the Twitter public API to collect data. There are a number of streaming API options that Twitter provides. We use a combination of the 1% random public feed stream and a keyword stream that uses health keywords. This combination allows us to increase data volumes. I think we could do a lot more with firehose access if we had it.

2. Mark, Could you elaborate more on how you pick your tweet data sets, since naming conventions differ among groups?

Thanks Ego I'm not clear on this question. Can we get a clarification/elaboration of what is meant by "data sets" and "naming conventions?"

3. For both speakers, Given privacy issues: what do you think is the applicability of these techniques when applied to local or even state level analysis with lower populations? What will be available, usable at 250,000 population or less of a city or say counties less than 30,000.

I mostly addressed this today. We will clearly hit a limit when we get to small populations or less significant health trends. We haven't hit those limits yet, but we're exploring them. I don't want to speculate what is or isn't possible until we try it.

4. How do you pull location sensitive data on public health from Twitter for local use?

Via Twitter I talked about this, but I'll add a citation to a paper we just published on this topic. It should be available freely online in the next month.

Mark Dredze, Michael J Paul, Shane Bergsma, Hieu Tran. Carmen: A Twitter Geolocation System with Applications to Public Health. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI), 2013.