the WINNOWER

REDDIT

PLOS Science Wednesday: Hi reddit, my name is Samuel Kou and I developed a model based on Internet search data that can track the spread of dengue fever – Ask Me Anything!

PLOSSCIENCEWEDNESDAY R/SCIENCE

Hi Reddit,
My name is Samuel Kou and I am a Professor of Statistics at Harvard University. My research interests include infectious disease tracking and forecasting, big data analytics, mathematical modeling in biology, and development of statistical methodologies.
I recently published an article titled Advances in using Internet searches to track dengue in PLOS Computational Biology. In the article, we presented a mathematical model that uses Google search data and government-provided clinical data to track dengue fever. The accurate tracking of dengue fever by our model in multiple countries shows that Internet search information, properly utilized, can help governments and health officials track infectious diseases, which is particularly important for countries with less advanced clinical based surveillance systems.
I will be answering your questions at 1pm ET. Ask me Anything!

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

DATE RECEIVED:
August 03, 2017

DOI:
10.15200/winn.150167.78271

ARCHIVED:
August 02, 2017

CITATION:
PLOSScienceWednesday , r/Science , PLOS Science Wednesday: Hi reddit, my name is Samuel Kou and I developed a model based on Internet search data that can track the spread of dengue fever – Ask Me Anything!, *The Winnower* 4:e150167.78271 , 2017 , DOI: 10.15200/winn.150167.78271

Can a frantic mother distort the data? When my daughter got dengue fever I had never really known much about it. I googled it obsessively for days: at work, at home, at my sisters. Since I was far away it was all I could do. I wasn't even on the same continent as her.

LittleRenay

This is a very good question. If the disease that one is tracking has a very small case number, a frantic mother's search can in theory distort the result. However, first, dengue fever that we are tracking has a big case number (often in tens of thousands). A frantic mother's search would hardly make any dent. Second, we are using multiple search query terms to estimate dengue occurrences. A frantic mother has to hit all the relevant search terms. Third, our method, ARGO, has the feature of automatically selecting the most useful Google search queries for estimation, and this makes it very robust against fluctuations in people's search queries.

What's you favorite programming language and what are he libraries you use the most?

Is your work open source? If so, could you link it? If not, why not?

Thanks for doing this AMA!

helpercolumn

My favorite programming language is R. It is free to use and has many packages publicly available at

CRAN. I use the time series R package xts the most. Our work is open source. The R package that implements our model, ARGO, can be downloaded from my web page: http://www.people.fas.harvard.edu/~skou/publication.htm

Since dengue fever is a mosquito-borne illness, could tracking precipitation or weather help improve the model?

shiruken

Yes, incorporating the weather information can potentially improve the model tracking. It is very interesting to explore this direction.

Do you have any advice for someone who has a background in molecular biology but is interested in switching into more computational biology? With the surge in using statistical/data science methods, I'm personally drawn towards projects using those methods, but have trouble getting on hands experience without returning to school for a more formal education (outside of things such as Coursera series).

aaaa29

I think reading some good textbooks and starting playing (meaning applying the methods) could be a good starting point. I find that one learns the best if he/she can directly apply the methods to a concrete data problem that he/she is really interested in.

Dengue Fever looks to be most widely transmitted in parts of the world that are relatively poor / without consistent access to the internet. (healthmap.org).

Is this model limited to certain geographies? If so, when would this be ready for prime time in places like South/ Central America or the Philippines? If not, how do you control for that lack of internet traffic in estimating the spread (ELI5)?

chargedanddangerous

We applied our method ARGO in five countries/states around the world: Mexico, Brazil, Thailand, Singapore and Taiwan. These places were chosen (i) due to the data availability and (ii) to explore the applicability of our approach in a diverse set of ecological situations, where dengue has been identified as an important local threat. Our method works best in countries with moderate to high Internet penetration and with sustained seasonal disease pattern. If a place has essentially no Internet connection, then one cannot use Google search information to track diseases there. Fortunately, more and more people are having Internet access now.

my name is Samuel Kou and I [...] can track the spread of dengue fever

How fast is dengue fever spreading?

LifeWin

Dengue is today one of the fastest-growing and most important mosquito-borne viral diseases in the world, with an estimated 390 million infections each year and threatening an estimated 3.9 billion people in 128 countries. Take Taiwan as an example. It had experienced very few cases before 2014, but then it was hit by big infection waves in 2014 and 2015.

Do you have a background in biostatistics or bioinformatics? And do you work directly with researchers in the microbiology field? I'm a summer studeny in a lab researching RSV but I have a friend who is a summer student working in a cancer research lab as their computer guy and it's incredibly interesting to me to see the mesh of technology and nature.

neofinger

I am a statistician. I collaborate with biologists, chemists, biophysicists. I do feel that being a statistician, one has many opportunities to work with people in multiple fields.

I would like to put together a data analytics platform but I don't have any of the knowledge on how to do it. What path do you suggest I take to make it come to life? Thanks!

Sooner_Nshn

I assume you are talking about a career path in data analytics. I think one can start from reading some good textbooks and trying to applying the (statistical or machine learning) methods. I find that one learns the best if he/she can directly apply the methods to a concrete data problem that he/she is really interested in.

Infectious disease researcher here. How well did your data align with local data about infectious outbreaks over the period monitored? Was it able to identify areas of infection before the infection was noted clinically? How sensitive is the model to outside interference, like this Reddit AMA, in artificially inflating the number of Dengue searches over a given period?

Arboviruses are notoriously bad in poor, low-infrastructure countries. I get the clinical and epidemiological reasoning here for going with dengue, but wouldn't you have been better served with a virus that gets more exposure in nations with the infrastructure to provide across the board internet access? I feel like there's a lot to control for with dengue in poor countries.

jquiz1852

Our model did accurate tracking of dengue in four countries: Brazil, Mexico, Thailand and Singapore. It did less well in Taiwan, possibly because Taiwan experienced less-consistent seasonal disease pattern from year to year. Governments traditionally rely on hospital-based reporting to track dengue; it is often lagged and limited with frequent post-hoc revisions. The wide availability of Internet throughout the globe provides the potential for an alternative way to track infectious diseases including dengue. This way of using Internet users' activity pattern to track dengue is particularly useful for less developed countries, where traditional hospital-based surveillance systems are often delayed or missing. This is what motivate us to study the tracking of dengue fever using Internet search information.

How do you validate the results from Google Trends and such tools? I've had reviewers complaining that I couldn't use these tools because of the normalization of data and unavailability of raw data.

Alopurinol

Yes, we did robustness studies to check how variation in Google Trends would affect our result. We found that our method ARGO is very robust again such variation. I think Google Trends data are quite useful but proper and rigorous statistical methods are needed.

Hi, will PR/news stories of this (and for example this reddit post) not interfere with the model and give a slightly exaggerated result, as more people might do a search on the topic after reading this post?

JNator

This is a good observation. Our method ARGO has a self-correction feature. Let me take a similar example to illustrate. Suppose our tracking alerts a government to take action. For instance, the introduction of an intervention to curb dengue activity, such as vector control or behavioral education, leads to a reduction in dengue cases. This may lead our model to temporarily over-predict incidence. However, once such an intervention has been established and remains active, our model will self-correct over time to predict the new levels of dengue activity.

As a mathematician what drew your interest in health and specifically epidemiology?

Sobjack

I am a statistician, so I play with data all the time. I am interested in applying statistical/mathematical models for real world problems. In this case, I am applying statistics in disease tracking.

To me this seems like common sense. If people in a region start searching for bomb making, is suspect a rise in terrorism there. If they search for bleeding from the eyes, I think the WHO and CDC should be alerted.

Why isn't this common practice?

wrek

Yes, the intuition is there. However, to properly utilize the information in Internet search, one has to be careful in building a mathematical model. For example, Google Flu Trends (GFT) is one of the earliest attempts aimed at tracking flu activity in the US and other countries using flu-related Google search activity in the population. Unfortunately, GFT was shown to poorly track flu activity in multiple occasions and was discontinued in 2015. In other words, proper and rigorous statistical methods are needed for the potential to be realized.

What made you interested in this field? It's not exactly a class we take in high school so where did you get your first exposure to bioinformatics?

Busamn

I am a statistician, so I play with data all the time. I am always interested in data and in building statistical/mathematical models. In computational biology, there are lots of data and lots of opportunities to build models. I was first exposed to bioinformatics when I was a graduate student. Back then bioinformatics is a new field. Nowadays, bioinformatics is well covered in the college courses.

Hasn't this methodology been discredited after Google did the same with flu? Are you predicting apples with oranges? How do you control for data being skewed and for multiple languages?

mirlb

Google Flu Trends (GFT) was shown to poorly track flu activity in multiple occasions and was discontinued in 2015. The inaccuracy of GFT leads people to question the utility of Internet data. In a recent study published in the Proceedings of the National Academy of Sciences, our team introduced a methodology that gives in fact accurate tracking of flu in the USA at the national level using Internet search data. In this article we extended our methodology to accurate tracking of dengue. What our works show is that (i) the wide availability of Internet throughout the globe provides the potential for an alternative way to reliably track infectious diseases, such as dengue, faster than traditional clinical-based systems; (ii) such a development would be particularly useful for less developed countries, but PROPER and RIGOROUS statistical methods are needed for the potential to be realized.

How does this translate to other infectious diseases, especially those that have similar symptoms?

Ldub52

Our method could also be used to track and map other infectious and mosquito-borne diseases, like Zika, malaria, yellow fever or Chikungunya. More studies are needed to explore these extensions.

How do you account for people who do searches on dengue fever just because they hear about it on the news?

Or to phrase it differently, how do you differentiate between people searching for information because some one they know is sick, from people searching simple because they've heard about it and want to stay informed?

DoomsdayDilettante

The intuition is that as long as there is a reasonably high correlation (within a moderate time window) between the Internet search volume and the disease counts, one can use mathematical models to take advantage of that correlation for prediction.

Do you see this model working for pathologies more common in the US, say Chlamydia or gonorrhea? Or even something like Zika?

Busamn

The method could also be used to track and map other infectious and mosquito-borne diseases, like Zika, malaria, yellow fever or Chikungunya. More studies are needed to investigate these possible extensions.

I live in The Yucatán where Dengue, Chinkunguña and Sika are lords of the lands. But I wonder ¿where in the world is Dengue more frequent?

Jolex41

Dengue is endemic throughout the tropical and subtropical world. It is today one of the fastest-growing and most important mosquito borne viral diseases in the world.

1. Is there a particular reason why you chose to do this for dengue?

2. A lot of the people who contact dengue in a country like Sri Lanka will usually live in remote rural areas- and while a lot of people have do have access to the internet, it won't be their first instinct to search Google up on their symptoms. Based off this, how do you know your data will be as accurate is recording the spread?

Thanks for doing this btw! I was personally taken down with dengue back in 2013, almost died. A lot of relatives and friends have also been personally affected by it in Sri Lanka, so I'm really glad to be seeing developments that are being made to tackle it as I really do feel like not enough is being done considering how serious and fatal is usually is.

Thisath

Thank you. Dengue is today one of the fastest-growing and most important mosquito-borne viral diseases in the world. Governments traditionally rely on hospital-based reporting, a method that is often lagged and limited with frequent post-hoc revisions, due to communication inefficiencies across local and national agencies and the time needed to aggregate information from the clinical to the state level. The wide availability of Internet throughout the globe provides the potential for an alternative way to track infectious diseases including dengue. This way of using Internet users' activity pattern to track dengue is particularly useful for less developed countries, where traditional hospital-based surveillance systems are often delayed or missing. This is what motivate us to study the tracking of dengue fever using Internet search information. Our method works best in countries with high Internet penetration and with sustained seasonal disease pattern. If a place has essentially no Internet connection, then one cannot use Google search information to track diseases there.

Hi there. I am interested in tracking outbreaks of multidrug resistant bacterial infections. Do you think your methodology could be applied to this?

Cycad

Yes, I think it can be potentially applied there.

For this type of research would you recommend the statistics master or the biostatistics master? Is a phd necessary to be succesful? Thank you very much for your time :)

RainyZRH

It doesn't matter if it is statistics master or biostatistics master. One of the coauthors is fresh college graduate, so a Ph. D. is good, but not that essential.

You are not tracking tengu fever. You are just tracking tengu fever inquiry.

Ikuyasu

We are using Google search information to provide near real-time estimate of dengue case counts in multiple countries.

Hello there, thank you for your work and for doing this AMA

Do you think it would be feasible to tweak your model and apply it to other fields?

thank you in advance

kerato

Yes, I think so. Our model was originally introduced to track flu in the US, and we extend it in this work to track dengue in multiple countries.