MÔ HÌNH DỰ BÁO SỚM DỊCH SỐT XUẤT HUYẾT DỰA VÀO GOOGLE TRENDS TẠI THÀNH PHỐ HỒ CHÍ MINH

Trần Ngọc Đăng[1], Lê Vĩnh Phát[2]

[1]Bộ môn SKMT, Khoa YTCC, Trường Đại học Y dược TP.HCM

[2]Trường Đại học Y Dược TP.HCM

Sốt xuất huyết Dengue (SXHD) là một bệnh truyền nhiễm tác nhân do vi rút truyền qua côn trùng phổ biến nhất. Mục tiêu của nghiên cứu là sử dụng nguồn dữ liệu lưu lượng tìm kiếm Google Trends index (GTI) xây dựng thành một mô hình có khả năng dự báo sớm dịch sốt xuất huyết tại TP.HCM nhằm mục đích hỗ trợ cho công tác giám sát và phòng chống dịch ở khu vực được thêm hiệu quả. Sử dụng phương pháp so sánh tương quan để ước tính mối liên hệ giữa GTI tra cứu với cụm từ "sốt xuất huyết" và dữ liệu số mắc SXHD tại TP.HCM, sau đó xây dựng một số mô hình dự đoán bằng hồi quy quasi-Poisson kết hợp những phép điều chỉnh nhằm loại bỏ sự tự tương quan của số liệu. Nghiên cứu đã cho thấy GTI tương quan cao với số mắc sốt xuất huyết với r2 = 0,74 và mô hình cuối cùng được chọn có khả năng dự đoán dịch SXHD tốt với độ chính xác là 87%, độ nhạy là 92,3% và độ đặc hiệu là 87%. Mô hình dự báo của chúng tôi cho thấy nguồn dữ liệu Google Trends rất có tiềm năng trong việc theo dõi và kiểm soát dịch SXHD ở TP.HCM. Những nghiên cứu sâu hơn nữa nhằm đánh giá tính hiệu quả của mô hình trong bối cảnh thực tế cần được thực hiện trong tương lai.

Từ khóa: Google Trends, mô hình dự báo, hồi quy Poisson, sự tự tương quan, sốt xuất huyết, thành phố Hồ Chí Minh.

I. ĐẶT VẤN ĐỀ

Sốt xuất huyết Dengue (SXHD) hay thường gọi là sốt xuất huyết là một bệnh truyền nhiễm cấp tính. Vi rút Dengue được lây truyền chủ yếu bởi muỗi cái Aedes aegypti. Tỷ lệ mắc SXHD trên toàn cầu tăng lên đáng kể qua những thập kỷ gần đây, nó khiến khoảng một nửa dân số thế giới đang nằm trong tình trạng nguy hiểm và là nguyên nhân hàng đầu gây bệnh tật nghiêm trọng thậm chí tử vong ở trẻ nhỏ. SXHD được tìm thấy ở khắp các vùng khí hậu nhiệt đới và cận nhiệt đới trên toàn thế giới, chủ yếu tập trung ở thành thị và bán thành thị [1]. Việt Nam nằm trong vành đai nhiệt đới, địa hình tự nhiên phức tạp, đồng thời chịu ảnh hưởng gió mùa Đông Bắc nên khí hậu luôn thay đổi từng năm và từng vùng [2]. Điều đó tạo điều kiện thuận lợi cho véc tơ truyền SXHD thích nghi, biến đổi và phát triển khó kiểm soát. Việt Nam có tỉ lệ mắc SXHD khá cao trong khu vực, trong vòng 7 tháng đầu năm 2017 cả nước đã ghi nhận 80.555 trường hợp mắc SXHD với 22 trường hợp tử vong, trong đó số trường hợp nhập viện là 69.085 ca. So với cùng kỳ năm 2016 (51.742/17) số mắc tăng 33,5%, số tử vong tăng 05 ca [3]. Thành phố Hồ Chí Minh (TP.HCM) với diện tích nhỏ nhất trong 20 tỉnh phía nam nhưng lại có mật độ phân bố dân cư thuộc hàng cao nhất nước với 3.927 người/km [4], đặc điểm thời tiết đặc trưng của nhiệt đới như nhiệt độ nóng ẩm, độ ẩm cao, lượng mưa lớn và đặc biệt có một mùa mưa kéo dài 5 - 6 tháng [5], góp phần tạo điều kiện thuận lợi cho

Tác giả liên hệ: Trần Ngọc Đăng, Khoa YTCC, Đại học Y dược TP.HCM

Email: ngocdangytcc@gmail.com

Ngày nhận: 05/04/2019

Ngày được chấp nhận: 07/05/2019

véc tơ truyền bệnh SXHD. Thực trạng công tác giám sát bệnh truyền nhiễm ở Việt Nam theo Thông tư 54/2015/TT-BYT quy định việc tổng hợp và báo cáo hàng tuần lên tuyến trên trong vòng 24 - 48 giờ sau khi được chẩn đoán [6]. Tuy nhiên quy trình này thường mất ít nhất một tuần để tổng hợp dữ liệu giám sát và công bố các báo cáo liên quan, thêm vào đó là sự trì hoãn hay gián đoạn công việc ở các tuyến do một số nguyên nhân khách quan và chủ quan như thiếu nguồn lực, chính sách đãi ngộ, trang thiết bị cơ sở y tế,… nên công tác giám sát dịch bệnh chưa được linh hoạt. Do đó, cần có một mô hình dự báo sớm dịch SXHD ở Việt Nam nói chung và TP.HCM nói riêng để hỗ trợ công tác phát hiện và kiểm soát dịch bệnh. Cả thế giới đang bước sang một cuộc cách mạng công nghiệp 4.0 với sự gia tăng nhanh chóng trong việc tạo ra các bộ dữ liệu kỹ thuật số khổng lồ được tích lũy qua nhiều năm, hay còn gọi là dữ liệu lớn (Big Data). Trong lĩnh vực chăm sóc sức khỏe, việc khai thác và nghiên cứu những dữ liệu có sẵn Big Data để tìm ra chiến lược mới tốt hơn dần thu hút được nhiều sự chú ý. Cụ thể vào năm 2009, Big Data đã ghi điểm trong y học khi Google sử dụng dữ liệu Google Trends để phân tích và dự đoán xu hướng ảnh hưởng, hướng lan truyền của dịch cúm H1N1. Xu hướng mà Google rút ra từ những từ khóa tìm kiếm liên quan đến H1N1 được chứng minh rất sát với kết quả do hai hệ thống cảnh báo cúm là SentinelGP và HealthStat đưa ra [7]. Với mong muốn kết hợp nguồn dữ liệu Google Trend và công tác dự báo dịch, chúng tôi quyết định thực hiện đề tài này với mục đích xây dựng một mô hình có khả năng dự báo sớm dịch sốt xuất huyết tại TP.HCM dựa vào dữ liệu lưu lượng tìm kiếm Google Trends index (GTI) để có thể hỗ trợ cho công tác giám sát và phòng chống dịch ở khu vực được thêm hiệu quả.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

1. Thiết kế nghiên cứu

Nghiên cứu tương quan sinh thái (Ecological study)

2. Đối tượng:

Số liệu về số ca mắc SXHD hàng tuần được thu thập từ hệ thống giám sát bệnh truyền nhiễm của Trung tâm Y tế Dự phòng (TTYTDP) TP.HCM từ năm 2012-2016. Số liệu được lấy phải là số hiệu chỉnh cuối cùng và lưu giữ trong hệ thống, nhằm tránh những sai sót như ca đã thay đổi chẩn đoán, ca chưa xác định, ca trùng, thiếu ca.

Lưu lượng tìm kiếm trên internet với cụm từ "sốt xuất huyết" được trích xuất từ ứng dụng Google Trends theo tuần từ năm 2012-2016 tại TP.HCM (gọi tắt là Google Trends index - GTI), tải xuống từ nguồn dữ liệu mở tại trang *https://trends.google.com.* Cú pháp nhập ở ô tìm kiếm chính xác chỉnh tả cụm từ "sốt xuất huyết", khu vực địa lý là "Việt Nam/ Hồ Chí Minh", danh mục là "Tất cả danh mục" và định dạng tìm kiếm là "Tìm kiếm trên web". Google Trends thể hiện chỉ số thống kê theo phần trăm: giá trị lưu lượng cao nhất trong khoảng thời gian được chọn bằng 100%, thấp nhất bằng 0%, các giá trị còn lại được tính theo mốc này. Bên cạnh từ khóa tìm kiếm "sốt xuất huyết" vẫn còn nhiều từ khóa liên quan khác có thể cho ra kết quả thõa mãn nhu cầu mà người tìm kiếm đang cần: "bệnh dengue", "dengue", "sốt dengue". Tuy nhiên, so sánh lưu lượng tìm kiếm của cả 4 từ khóa với nhau trong cùng một khoảng thời gian từ năm 2012 - 2016 cho kết quả: từ khóa "sốt xuất huyết" chiếm con số áp đảo 82%. Bên cạnh đó cụm từ "sốt xuất huyết" còn là một danh từ thuần việt, đầy đủ nghĩa và không hạn chế đối tượng đọc hiểu. Do đó chúng tôi chỉ sử dụng từ khóa "sốt xuất huyết" để trích xuất dữ liệu trong nghiên cứu này.

3. Phương pháp

Bất kì dạng chuỗi dữ liệu theo thời gian nào đều thường xảy ra hiện tượng tự tương quan (Auto Correllation-AC), là một hiện tượng các thành phần của một chuỗi các quan sát theo thời gian hay không gian tự ảnh hưởng lên nhau. Nguyên nhân khách quan thường là do tính "quán tính" của số liệu, sự biến động của quan sát thứ i có thể tác động vào kỳ thứ i + k [8] (k được gọi là độ trễ lag của số liệu). Ở nghiên cứu của chúng tôi, biến độ trễ của số mắc SXHD được sử dụng để kiểm soát sự tự tương quan này. Dùng mô hình hồi quy Poisson (có hiệu chỉnh cho sự phân tán số liệu over-dispersion bằng quasi-Poisson) để xác định mối liên quan của tác động trễ (lag) của lưu lượng tìm kiếm GTI với số mắc SXHD sử dụng một số biến đổi để loại bỏ sự tự tương quan của biến SXHD

Mô hình chung được biểu diễn như sau:

Yt ~ quasi-poisson (μt)

$$\log \mu_t = \alpha + \beta_1 \text{ Lag GTI}_{t-k} + \beta_{AC} AC$$

$$= \text{Basis TSR} + \beta_{AC} AC$$

Trong đó:

Yt: Số ca mắc SXHD được dự đoán của tuần t

μt: Số mắc SXHD trung bình dự đoán của mô hình Poisson

Lag GTI$_{(t-k)}$: lưu lượng tìm kiếm GTI của tuần t với độ trễ k tuần (k = 0, 1, 2, 3)

α, β_1, β_{AC}: hệ số hồi quy

Basis TSR: Mô hình tiên lượng nền tảng

AC - Auto Correlation: sự tự tương quan của biến SXHD, lần lượt là phần dư của Yt-1 , Yt, logarit của (Yt-1+1)

Sau đó xây dựng mô hình tiên lượng dựa trên phân tích đơn biến số ca mắc SXHD và lưu lượng tìm kiếm GTI có sử dụng các phép biến đổi để loại bỏ sự tự tương quan của biến SXHD. Trong nghiên cứu này, tổng cộng chúng tôi phân thành 7 mô hình:

1. Basis TSR: mối liên quan tuyến tính giữa số ca mắc SXHD và độ trễ 1 tuần của lưu lượng tìm kiếm GTI (mô hình nền tảng).

2. Basis TSR + AC: Lag(Residuals,1): mối liên quan tuyến tính giữa số ca mắc SXHD và độ trễ 1 tuần của lưu lượng tìm kiếm GTI, loại bỏ sự tự tương quan của SXHD bằng độ trễ 1 tuần của phần dư mô hình nền tảng.

3. Basis TSR + AC: Lag(SXH,1): mối liên quan tuyến tính giữa số ca mắc SXHD và độ trễ 1 tuần của lưu lượng tìm kiếm GTI, loại bỏ sự tự tương quan của SXHD bằng độ trễ 1 tuần của số ca mắc SXHD.

4. Basis TSR + AC: Lag(log(SXH+1),1): mối liên quan tuyến tính giữa số ca mắc SXHD và độ trễ 1 tuần của lưu lượng tìm kiếm GTI, loại bỏ sự tự tương quan của SXHD bằng độ trễ 1 tuần của logarit số mắc SXHD cộng 1 (cộng 1 vào số mắc SXHD nhằm loại bỏ những dữ liệu bị giá trị 0).

5. TSR Lag(GTI,2) + AC: Lag(log(SXH+1),2): mối liên quan tuyến tính giữa số ca mắc SXHD và độ trễ 2 tuần của lưu lượng tìm kiếm GTI, loại bỏ sự tự tương quan của SXHD bằng độ trễ 2 tuần của logarit số mắc SXHD cộng 1.

6. TSR Lag(GTI,3) + AC: Lag(log(SXH+1),3): mối liên quan tuyến tính giữa số ca mắc SXHD và độ trễ 3 tuần của lưu lượng tìm kiếm GTI, loại bỏ sự tự tương quan của SXHD bằng độ trễ 3 tuần của logarit số mắc SXHD cộng 1.

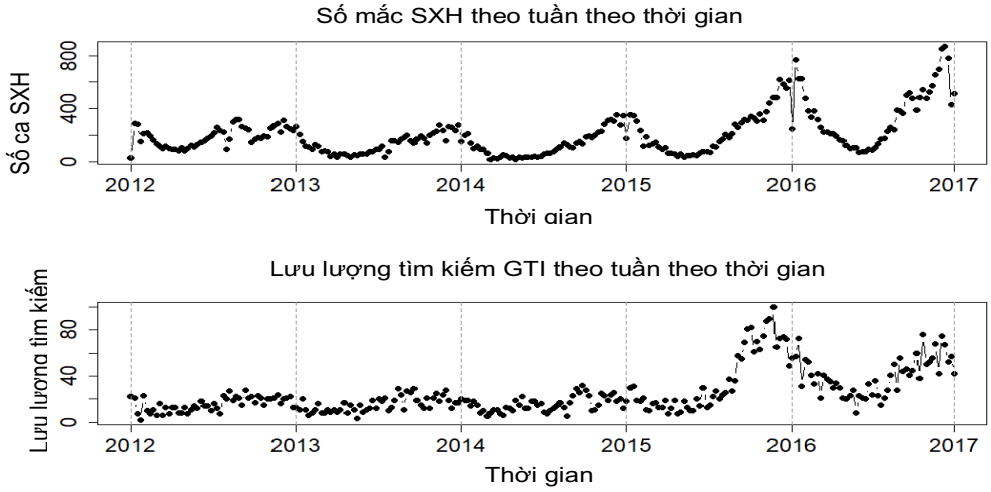
7. TSR Lag(GTI,0) + AC: Lag(log(SXH+1),1): mối liên quan tuyến tính giữa số ca mắc SXHD và lưu lượng tìm kiếm GTI, loại bỏ sự tự tương quan của SXHD bằng độ trễ 1 tuần của logarit số mắc SXHD cộng 1

Mô hình được chúng tôi lựa chọn để dự báo là mô hình cho chỉ số phân tán (dispersion)

thấp nhất trong tất cả các mô hình. Chỉ số phân tán càng thấp, mô hình dự báo càng tốt [9]. Sự có mặt của tự tương quan càng hạn chế bao nhiêu là một dấu hiệu cho thấy mô hình tiên lượng xây dựng được chuẩn và cho những ước lượng tốt bấy nhiêu. Thông thường các bài kiểm tra sự tự tương quan thường được xem như là kiểm tra sự sai lệch của mô hình. Việc loại bỏ hoặc giảm thiểu đến mức thấp nhất có thể sự tự tương quan sẽ chứng minh được mô hình được lựa chọn có tính dự báo cao. Biến tiên lượng lưu lượng tìm kiếm GTI được phân thành các nhóm tương ứng với các khoảng giá trị bách phân vị < 50, 50 - < 75, 75 - < 95 và ≥ 95. Sau đó mô hình dự báo đã chọn được biểu diễn thông qua các nhóm phân loại này. Hệ số ước tính β sau khi tính toán từ mô hình được chuẩn hóa và làm tròn thành điểm dự báo. Mô hình dự báo dịch được xây dựng với biến tiên lượng là biến định lượng lưu lượng tìm kiếm GTI, biến kết cuộc là biến nhị giá có hai giá trị là: có dịch và không dịch. Dựa theo một nghiên cứu cùng mục đích xây dựng nên mô hình dự báo dịch SXHD của Phùng Trí Dũng và các cộng sự ở Thành phố Cần Thơ, tuần được đánh giá là khả năng có dịch khi số ca mắc nằm trong khoảng ≥ 95% số ca mắc thực tế [10]. Đánh giá khả năng dự báo của mô hình dựa vào diện tích dưới đường cong ROC (Receiver Operating Characteristic), theo đó diện tích dưới đường cong (Area Under the Curve – AUC) càng lớn thì mô hình tiên lượng càng có khả năng dự báo tốt. Và một mô hình có AUC > 0.8 thì có thể được xem xét ứng dụng vào thực tiễn[11]. Sử dụng chỉ số Youden để tìm ra điểm cắt tối ưu. Chỉ số Youden chính là tổng giá trị của độ nhạy và độ đặc hiệu, do đó tại điểm cắt có chỉ số Youden cao nhất nghĩa là mô hình tại điểm cắt có độ nhạy và độ đặc hiệu tối ưu. Tiếp theo, chúng tôi tính phần trăm dự báo chính xác của mô hình tại điểm cắt tối ưu.

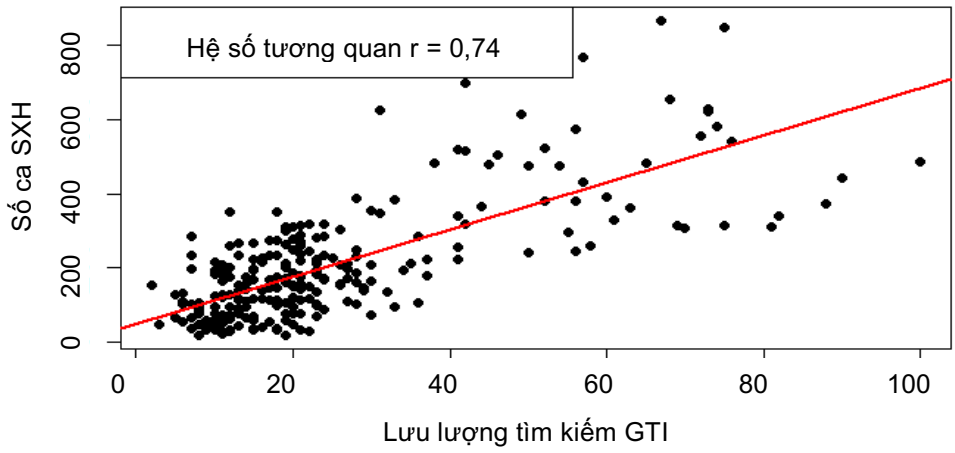
III. KẾT QUẢ

Từ năm 2012 - 2016, tại TP.HCM có tổng cộng 53.384 ca mắc SXHD được báo cáo. Số mắc thường tăng cao vào giữa năm và kéo dài đến cuối năm, không thấy tính chu kì đối với các năm xảy ra dịch lớn. Số mắc cao nhất trong khoảng nửa cuối 2015 đến đầu 2016. Biến số lưu lượng tìm kiếm GTI giữa năm 2015 đến đầu 2016 và nửa cuối năm 2016 có sự tăng đột biến, theo đó bật ra khỏi tính chu kỳ so với các năm trước (Biểu đồ 1).



Biểu đồ 1: Phân bố SXHD và GTI tại TP.HCM năm 2012 đến 2016

Lưu lượng tìm kiếm GTI với từ khóa "sốt xuất huyết" theo tuần có tương quan dương với số ca mắc SXHD tuần với hệ số tương quan r là 0,74 (KTC 95% 0,68 – 0,79), biểu diễn rõ qua đường thẳng màu đỏ - đường hồi quy tuyến tính mối tương quan giữa 2 biến (Biểu đồ 2).



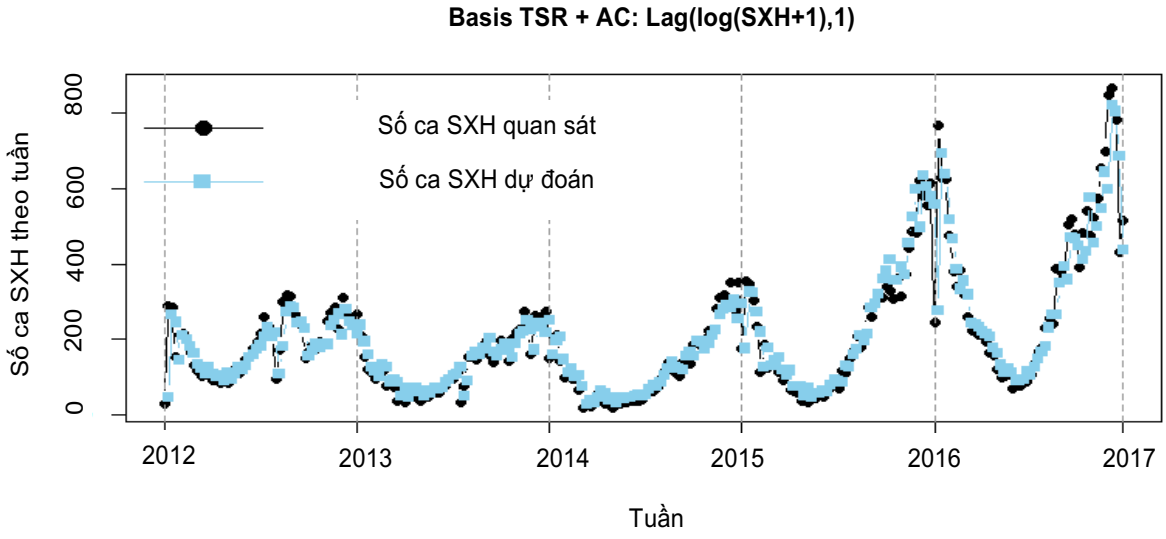
Biểu đồ 2. Phân tán đồ mối tương quan giữa SXHD và GTI

Từ đó cho thấy tỉ lệ biến thiên của lưu lượng tìm kiếm GTI có thể giải thích được 55% số ca mắc SXHD. Số ca mắc thực tế và số ca mắc dự đoán được mô tả thông qua 7 mô hình (Bảng 1), và mô hình thứ 4 với hệ số tương quan 0,92 (KTC 95% 0,9 – 0,94), hệ số phân tán 18,6 cho biểu đồ thể hiện đường dự đoán nắm bắt tốt nhất số ca mắc SXHD thực tế (Biểu đồ 3).

Bảng 1. Thống kê chỉ số phân tích của 7 mô hình

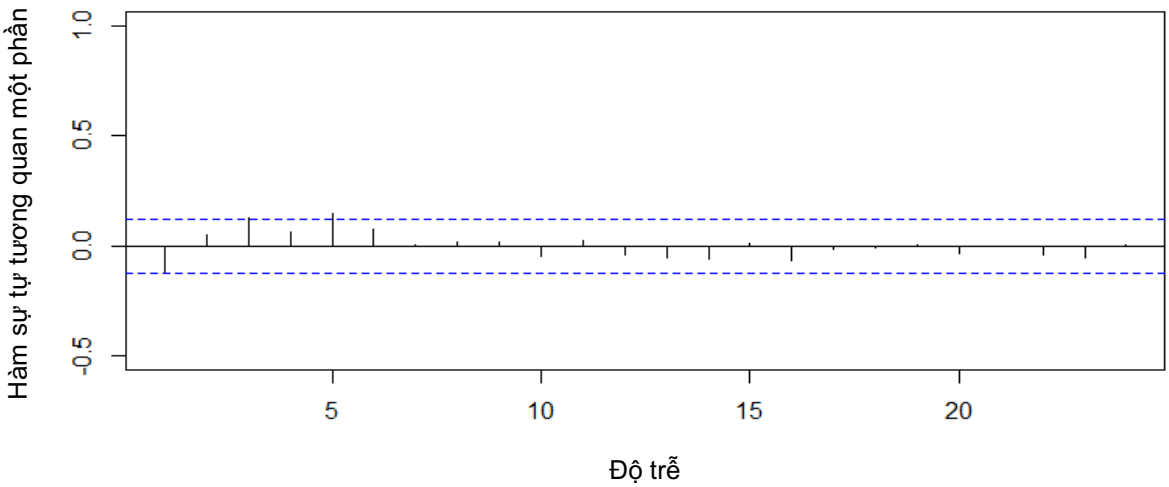
STT	Mô hình	Hệ số hồi quy	SE	Giá trị Z	Giá trị P	2,5%	97,5%	Độ phân tán
1	Basis TSR	0,023	0,001	18,374	<0,001	0,020	0,025	52,6
2	Basis TSR + AC:Lag(Residuals,1)	0,021	0,001	19,788	<0,001	0,019	0,023	32,2
3	Basis TSR + AC:Lag(SXH,1)	0,008	0,001	5,342	<0,001	0,005	0,011	28,5
4	Basis TSR + AC:Lag(log(SXH+1),1)	0,005	0,001	4,494	<0,001	0,003	0,007	18,6
5	TSR Lag(GTI,2) + AC:Lag(log(SXH+1),2)	0,006	0,001	5,049	<0,001	0,004	0,009	22,6

STT	Mô hình	Hệ số hồi quy	SE	Giá trị Z	Giá trị P	2,5%	97,5%	Độ phân tán
6	TSR Lag(GTI,3) + AC:Lag(log(SXH+1),3)	0,007	0,001	5,534	<0,001	0,005	0,009	24,4
7	TSR Lag(GTI,0) + AC:Lag(log(SXH+1),1)	0,004	0,001	3,839	<0,001	0,002	0,006	19,0



Biểu đồ 3. Số ca SXH quan sát và số ca SXH dự đoán từ mô hình tốt nhất (Mô hình 4)

Các giá trị phần dư của mô hình thứ 4 được thể hiện qua (Biểu đồ 4). Trong đó các giá trị phần dư phân bố đồng đều quanh đường số 0, chứng tỏ mô hình 4 dự đoán rất tốt số ca mắc SXH dựa vào dữ liệu GTI.



Biểu đồ 4. Phần dư của mô hình 4 theo các thời gian trễ (lag)

Biến tiên lượng lưu lượng tìm kiếm GTI được phân thành các nhóm tương ứng với các khoảng giá trị bách phân vị < 50, 50 – < 75, 75 – < 95 và ≥ 95. Sau đó mô hình dự báo đã chọn được biểu diễn thông qua các nhóm phân loại này. Hệ số ước tính β sau khi tính toán từ mô hình được chuẩn hóa và làm tròn thành điểm dự báo (Bảng 2).

Bảng 2. Thang điểm dự báo của mô hình

Phân nhóm	Hệ số hồi quy	KTC 95%	Điểm
GTI trễ 1 tuần (%)			
< 19	Nhóm nền		
≥ 19 - < 28	0,26	1,15 – 1,45	26
≥ 28 - < 69,1	0,5	1,45 – 1,88	50
≥ 69,1	0,56	1,46 – 2,08	56
Số mắc SXHD trễ 1 tuần (lấy logarit)			
< 5,123964	Nhóm nền		
≥ 5,123964 - < 5,585372	0,75	1,87 – 2,39	75
≥ 5,585372 - < 6,301490	1,01	2,4 – 3,14	101
≥ 6,301490	1,43	3,49 – 4,98	143

Dựa theo một nghiên cứu cùng mục đích là xây dựng mô hình dự báo dịch SXHD của Phùng Trí Dũng và các cộng sự tại Thành phố Cần Thơ và khu vực Đồng bằng sông Cửu Long, nghiên cứu đã chia số quan sát số mắc SXHD theo tuần thành hai nhóm: có dịch và không dịch. Theo đó, tuần được đánh giá là có dịch khi số ca mắc nằm trong khoảng ≥ 95% số ca mắc thực tế (≥ 543,7 ca). Chúng tôi kiểm tra độ chính xác của mô hình dự báo được chọn bằng đường cong ROC với kết quả diện tích vùng dưới đường cong ROC (Area Under the Curve) thu được là AUC = 0,969 với KTC 95% 0,936 – 1, cho thấy khả năng dự báo tốt của mô hình này. Mô hình dự báo tốt nhất ở điểm cắt 0,051 (với tổng điểm dự báo là 152), độ chính xác 87%, khi đó độ nhạy của mô hình dự báo là 0,923 và độ đặc hiệu là 0,87.

IV. BÀN LUẬN

Nghiên cứu của chúng tôi đã tìm thấy mối liên quan giữa lưu lượng tìm kiếm trên internet với cụm từ "sốt xuất huyết" và số mắc SXHD thực tế với hệ số tương quan là 0,74 (KTC 95% = 0,68 - 0,79). Theo đó, mô hình tốt nhất có thể dự báo sớm dịch SXHD một tuần với độ chính xác là 87% (chỉ số AUC = 0,969). Một số nghiên cứu trước cũng sử dụng lưu lượng tìm kiếm GTI và đã xây dựng thành công các mô hình dự báo dịch bệnh ở Mỹ, Bangkok, Singapore, Mexico và Trung Quốc [12 - 18]. Ví dụ, nghiên cứu tại Singapore [16] sử dụng GTI để dự báo sốt xuất huyết với hệ hố tương quan 0.931 và giá trị AUC = 0.906. Tuy nhiên một nhược điểm của các nghiên cứu trước là sử dụng mô hình thống kê rất phức tạp do đó khó ứng dụng trong thực tế. Nghiên cứu của chúng tôi đã chuyển dịch mô hình thống kê sang dạng

điểm số rất đơn giản, do đó gia tăng khả năng ứng dụng mô hình vào công tác giám sát và phòng chống dịch bệnh thường quy.

Mô hình tốt nhất của chúng tôi ứng dụng thực tế có thể dự báo sớm được dịch SXHD sẽ xảy ra trong tương lai 1 tuần. Vì tác nhân và cơ chế lây truyền giống nhau nên mô hình dự báo của chúng tôi có thể được sử dụng để theo dõi bệnh do các vi rút Arbo (nhóm A và B) thường gặp khác như viêm não Nhật Bản, Zika. Một nghiên cứu khác của cùng nhóm tác giả đã thành công xây dựng mô hình dự báo dịch Zika tại Brazil và Columbia sử dụng GTI [19], tuy nhiên một nghiên cứu tương tự cho Việt Nam chưa thể áp dụng vì số ca mắc Zika của Việt Nam khá thấp.

Nghiên cứu này có hai điểm hạn chế chính: Thứ nhất, mô hình của chúng tôi không thể áp dụng được ở tuyến xã phường, hoặc quận huyện do sự hạn chế về độ phân giải không gian của dữ liệu. Hiện nay dữ liệu GTI chỉ cung cấp lưu lượng tìm kiếm cho toàn thành phố Hồ Chí Minh. Trong tương lai nếu có thể thu thập được dữ liệu có độ phân giải không gian tốt hơn và kết hợp với hệ thống thông tin địa lý (Geographic Information System – GIS) thì sẽ cải thiện được mô hình rất nhiều. Thứ hai, nghiên cứu này chỉ mang tính chất "thăm dò", cần phải đánh giá lại mô hình trong bối cảnh thực tiễn trước khi thực sự áp dụng vào công tác dự báo dịch thường quy.

V. KẾT LUẬN

Mô hình dự báo của chúng tôi cho thấy nguồn dữ liệu Google Trends rất có tiềm năng trong việc theo dõi và kiểm soát dịch SXHD ở TP.HCM. Những nghiên cứu sâu hơn nữa nhằm đánh giá tính hiệu quả của mô hình trong bối cảnh thực tế cần được thực hiện trong tương lai.

Lời cảm ơn

Nhóm nghiên cứu xin chân thành cám ơn ông Nguyễn Trí Dũng, giám đốc TTYTDP TP.HCM và tập thể cán bộ Khoa Kiểm soát bệnh truyền nhiễm đã hỗ trợ trong quá trình thực hiện nghiên cứu.

TÀI LIỆU THAM KHẢO

1. **World Health Organization (2017)** Fact sheet: Dengue and severe dengue, http://www.who.int/mediacentre/factsheets/fs117/en/.

2. **Trần Mạnh Thường (2005),** Việt Nam - Văn hóa và Du lịch, NXB Thông Tấn, 15 - 16.

3. **Bộ Y Tế (2017),** Cục Y tế dự phòng báo cáo tình hình Sốt xuất huyết lên Bộ Y tế, Government Document, 16.

4. **Cục Thống kê TP.HCM (2015),** Niên giám Thống kê về Dân số và Lao động, Government Document, 19.

5. **Viện Vệ sinh Dịch tễ Trung Ương (NIHE) (2013)** ,Hỏi đáp về dịch bệnh Sốt xuất huyết, http://nihe.org.vn/vn/tin-tuc-su-kien/giam-sat-va-phong-chong-dich-benh/hoi-dap-ve-dich-benh-sot-xuat-huyet/mua-mua-mua-cua-sot-xuat-huyet-c12320i14654.html.

6. **Bộ Y Tế (2015)** Thông tư 54 "Hướng dẫn chế độ thông tin báo cáo và khai báo bệnh, dịch bệnh truyền nhiễm", Government Document, 51.

7. **Cook S, Conrad C, Fowlkes AL, Mohebbi MH (2011),** "Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic". *PloS one,* **6(8),** e23610, 52.

8. **Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, et al (2013),** "Time series regression studies in environmental epidemiology". International journal of epidemiology, **42(4),** 1187 - 1195, 79.

9. Imai C., Armstrong B., Chalabi Z., Mangtani P et al (2015), "Time series regression model for infectious disease and weather". *Environmental research*, **142**, 319 - 327.

10. Phung D., Cunrui H., Shannon R., Cordia C et al (2015), "Identification of the prediction model for dengue incidence in Can Tho city, a Mekong Delta area in Vietnam". *Acta tropica,* **141**, 88 - 96.

11. Pencina M.J., D'Agostino R.B., Vasan R.S. (2008), "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond". *Statistics in medicine,* **27(2)**, 157 - 172.

12. Dugas A.F., Jalalpour M., Gel Y., Levin S et al. (2013), "Influenza forecasting with Google flu trends". *PloS one,* **8(2),** e56176.

13. Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L et al (2009), "Detecting influenza epidemics using search engine query data". *Nature*, **457(7232)**, 1012.

14. Gluskin R.T., Johansson M.A., Santillana M., Brownstein J.S. (2014), "Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends ". *PLOS Neglected Tropical Diseases,* 84.

15. Gu Y., Chen F., Liu T., Lv X et al (2015) , "Early detection of an epidemic erythromelalgia outbreak using Baidu search data". *Scientific reports,* **5**, 12649.

16. Althouse B.M., Ng Y.Y., Cummings D.A.T. (2011), "Prediction of dengue incidence using search query surveillance". *PLoS neglected tropical diseases,* **5(8)**, e1258.

17. Xu Q., Gel Y.R., Ramirez L.R., Nezafati K et al (2017) "Forecasting influenza in Hong Kong with Google search queries and statistical model fusion". *PloS one,* **12(5)**, e0176690.

18. Yuan Q., Nsoesie E.O., Lv B., Peng G et al (2013) , "Monitoring influenza epidemics in china with search query from baidu". *PloS one,* **8(5)**, e64323.

19. Morsy S., Dang T.N., Kamel M.G., Zayan A.H et al (2018): Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends. *Epidemiology & Infection*, **146(13)**: 1625 - 1627.

Summary

A PREDICTION MODEL FOR DENGUE OUTBREAKS IN HO CHI MINH CITY BY USING GOOGLE TRENDS

Dengue Fever (DF) is the most common viral disease transmitted by an insect in the world. The objective of this study was to use Google Trends Index (GTI) to build a prediction model for Dengue outbreaks in HCMC to support the control and prevention activities in local area. A correlation was measured to estimate the association between GTI search with the phrase “sốt xuất huyết” and dengue surveillance data in HCMC, which was used to build prediction models by using quasi-Poisson regression with some adjustment to eliminate the auto-correlation of the data. GTI correlated highly with dengue data with r2 = 0.74 and our final prediction model had a good prediction power with the accuracy of 87%, the sensitivity of 92.3% and specificity of 87%. Our prediction model shows some potential use in monitoring and controlling the dengue outbreak in Ho Chi Minh City. The further study is warranted to test the efficacy of the model in a real context.

Keywords: Google Trends, prediction model, Poisson, auto-correlation, Dengue, Ho Chi Minh City.