Correspondence

A look back: investigating Google Flu Trends during the influenza A(H1N1)pdm09 pandemic in Canada, 2009–2010

To the Editor

Recently, my colleagues and I found Google Flu Trends (GFT) to effectively estimate the proportion of sentinel physician visits related to influenza-like illness (ILI) reported by the Public Health Agency of Canada (PHAC) on a national level in Canada in 2010–2014 [1]. However, we omitted the 2009 H1N1 pandemic period from our analysis as we were uncertain about retrospective revisions to GFT estimates in Canada [1]. In the United States, GFT underestimated traditional surveillance values during the first pandemic wave [2]. Google then changed its US GFT model in September 2009 [3] and, using this new model, more accurately represented the second wave of the 2009 H1N1 pandemic in the United States [2, 3]. Olson *et al.* and Cook *et al.* [2, 3] describe and compare revised and original US GFT estimates during the pandemic period, helping to establish a record of the real-time performance of these estimates in the United States; however, similar analyses are unavailable for Canada. GFT performance during the 2009 H1N1 pandemic has also been described in other countries [4–7]. In Canada, GFT estimates during the 2009 H1N1 pandemic have been examined on a provincial level in Manitoba [8, 9]; however, to my knowledge, they have not been examined nationally in this country during that time. Although beginning in August 2015, Google stopped posting real-time GFT estimates online, GFT estimates are still available to some researchers [10] and previous estimates remain publicly available [11]. Documentation of the accuracy of GFT estimates during the 2009 H1N1

pandemic period can inform future use and interpretation of these data. This letter extends our previous analysis [1] to investigate retrospective revisions to GFT estimates during the 2009 H1N1 pandemic in Canada and compares GFT estimates to ILI consultation rates reported by PHAC during this time.

I accessed GFT estimates for Canada [12, 13] using the Internet Archive [14], which has been used by others [15] to find previously available GFT estimates. To determine when GFT was introduced in Canada and cross-reference dates, I used The Official google.org blog [16] and news reports. For Canada, GFT estimates are interpreted as 'ILI cases per 100 000 physician visits' [17]. Similar to our previous analyses [1], I converted GFT estimates to percentages (%GFT) and obtained archived FluWatch reports from PHAC, from which I manually entered ILI consultation rates [18] and converted these to percentages (%PHAC); I assessed how well GFT estimated %PHAC by comparing the magnitude and timing of peaks in %GFT to those in %PHAC and by calculating Spearman correlation coefficients between %GFT and %PHAC, which is consistent with metrics used by others [2, 4]. I included weekly data for 24–30 August 2008 to 5–11 September 2010 to include the entire 2008–2009 influenza season and allow an overlap of two full weeks with our previous analysis, which began the week of 29 August 2010 [1]; this enabled documentation of differences between archived GFT estimates and those included in our previous work. I defined the pandemic waves as 12 April–29 August 2009 (wave 1) and 30 August 2009–30 January 2010 (wave 2), based on definitions used by PHAC [19]. Ethics approval was not required for the use of these publicly available data. Analyses were conducted using SAS v. 9.4 (SAS Institute Inc., USA) and R v. 3.2.5 [20].

GFT estimates for Canada became available 8 October 2009 [21], which was documented in a Google blog post of the same date entitled 'Google
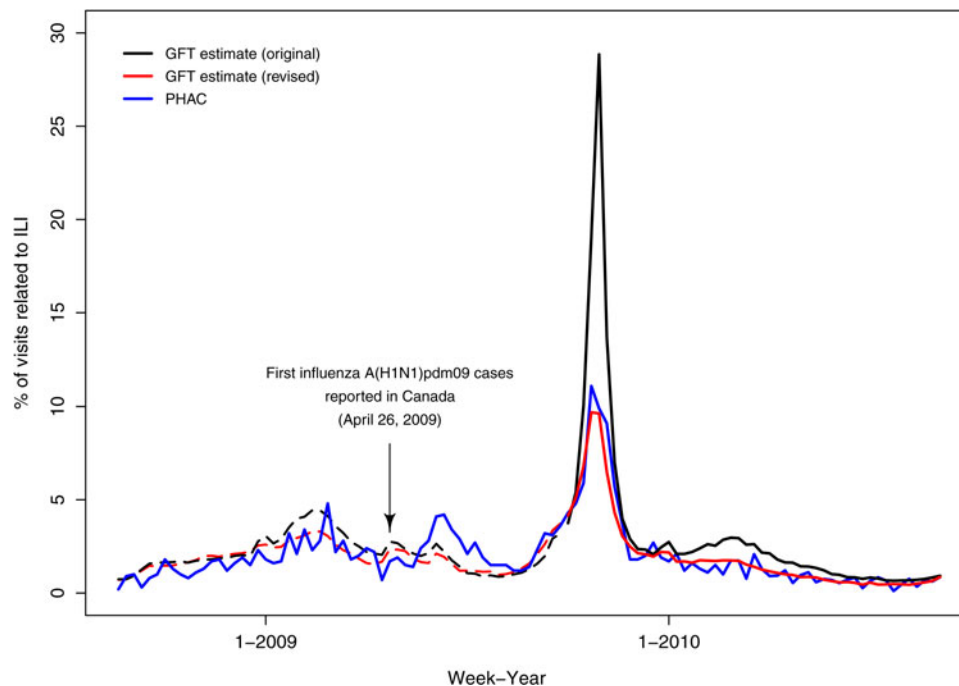
Fig. 1. Comparing Google Flu Trends (GFT) estimates to Public Health Agency of Canada (PHAC) influenza-like illness (ILI) consultation rates during the influenza seasons affected by the H1N1 pandemic, 24–30 August 2008 to 5–11 September 2010. Dashed lines indicate estimates for the period before GFT was introduced in Canada. The first influenza A(H1N1)pdm09 cases were reported in Canada on 26 April 2009 [26].

Flu Trends expands to 16 additional countries'; however, these countries were not named [22]. These estimates were made available retrospectively back to 28 September 2003 [13]. Then, some time between 15 September and 31 December 2010, Canadian GFT estimates were revised and replaced. From the Internet Archive, GFT estimates available for Canada on 31 December 2010 ('revised' estimates) [23] differed from those available on 15 September 2010 ('original' estimates) [24]. This corresponds to a google.org blog post-dated 12 November 2010 stating that Google was 'refreshing ... models in 13 countries' [25] and, although the countries affected were not specified, based on the present analysis, this included Canada. Therefore, based on these findings, our previous analyses [1] included revised %GFT estimates from 29 August 2010 until this update (an estimated 11 weeks) that we thought were prospectively estimated, but were actually retrospectively estimated. For the 2-week overlap between the two studies that I have included in this analysis, original and revised estimates were similar (absolute difference = 0·1–0·2 percentage points).

During the first wave of the 2009 H1N1 pandemic, both original and revised %GFT estimates had two similarly sized peaks, with the original %GFT estimates peaking slightly higher, reaching maxima of 2·8% in week 17 (26 April–2 May 2009) and 2·6% in week 23 (7–13 June 2009) (Fig. 1). The first of these peaks in %GFT was coincident with the reporting of the first pH1N1 cases in Canada on 26 April 2009 [26]. This could be a response to possible increased search queries during this time, or may correspond to a true increase in healthcare use, as a slight increase in %PHAC is also observable during this and the following week (Fig. 1). The second of these peaks was coincident with the maximum peak in %PHAC during wave 1; however, both the original and revised %GFT estimates *underestimated* this second, larger peak in %PHAC during this first wave in weeks 23–24 by 37% (original %GFT) and 50% (revised %GFT) (Fig. 1). Original and revised %GFT estimates showed little correlation with %PHAC during the first pandemic wave ($\rho = 0\cdot29$, $P = 0\cdot22$ and $\rho = 0\cdot21$, $P = 0\cdot37$, respectively).

During the second wave of H1N1, although original %GFT estimates correlated with %PHAC ($\rho = 0\cdot79$, $P < 0\cdot0001$), they *overestimated* the magnitude of the %PHAC peak by 160%, reaching a maximum of 29% compared to a maximum of 11% for %PHAC, and peaked 1 week later (Fig. 1). In comparison, revised

%GFT estimates were more strongly correlated with % PHAC ($\rho = 0.90$, $P < 0.0001$) and much closer in magnitude to %PHAC values, peaking 13% lower (9·7% vs. 11%) and during the same week (Fig. 1) as %PHAC data.

Similar to these findings for Canada, on a national level in the United States, GFT underestimated the peak in traditional surveillance values during the first wave of the 2009 H1N1 pandemic [2]. However, in contrast to the situation for Canada, in the United States, Google began prospectively estimating revised GFT estimates, in real time, before the highest peak in traditional surveillance values occurred during the second pandemic wave in that country [3]. These revised estimates were highly correlated with traditional surveillance values and more accurately represented the remainder of this second wave of the pandemic on a national level in the United States [2, 3]. In contrast, in Canada, more accurate, revised GFT estimates were not available until after the pandemic had ended. In Europe, the performance of GFT varied by country; however, similar to the present study, large overestimates of peak magnitude during the second pandemic wave were also observed, with absolute differences between GFT estimates and traditional surveillance values being greatest for France and Hungary [4].

It is advantageous to retrospectively revise models based on new data in an effort to improve them for future use. However, such revisions should be clear and well-documented so that resulting data can be appropriately interpreted and the prospectively achieved success of the model can be realistically assessed. At least two Canadian provinces had incorporated GFT estimates into their influenza surveillance reports before these estimates went offline [27, 28]; however, the impact of the loss of real-time, publicly available GFT estimates in Canada depends on if and how these estimates were previously used. If access to and examination of current GFT estimates for Canada is considered, the revisions outlined herein may facilitate interpretation, especially in considering pandemic scenarios.

This study has limitations. I manually entered ILI consultation rates from FluWatch reports and examined bar charts; possible post-reporting updates would not necessarily be incorporated in this analysis. However, updated %PHAC peak values during the first and second pandemic waves were similar to the values included in this analysis: in weeks 23, 24, and 43, there was a difference of −1·5 to 1·3 ILI-related visits/1000 physician visits, with %PHAC peaking

during the first wave in week 23 (FluWatch, 1 December 2015, personal communication), the same week %GFT peaked. Furthermore, archived GFT data were only available at certain time points; therefore, not all original estimates are publicly available. Only national ILI consultation rates are publicly provided by PHAC; therefore, examination of provincial data was beyond the scope of this analysis. National patterns reported here may not represent what would have been seen provincially.

In summary, GFT estimates for Canada were not available until the beginning of the second wave of the 2009 H1N1 pandemic. Currently available GFT estimates were retrospectively revised and do not represent what would have been available in real-time during the pandemic. During the first pandemic wave, both original and revised %GFT estimates underestimated peak ILI rates reported by PHAC; however, neither of these estimates were available in real-time. During the second pandemic wave, original GFT estimates became available; although well-correlated with %PHAC, these original GFT estimates overestimated the magnitude of the %PHAC peak during this second pandemic wave by 160%. Revised GFT estimates better estimated %PHAC during the second pandemic wave than original GFT estimates; however, these revised estimates only became available after the pandemic had ended. These results show how GFT estimated traditional surveillance data nationally in Canada in real-time during the most recent pandemic and supplement the historic record provided by Google.

Declaration of Interest

None.

References

1. **Martin LJ, Lee BE, Yasui Y.** Google Flu Trends in Canada: a comparison of digital disease surveillance data with physician consultations and respiratory virus surveillance data, 2010–2014. *Epidemiology and Infection* 2016; **144**: 325–332.

2. **Olson DR, et al.** Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology* 2013; **9**: e1003256.

3. **Cook S, et al.** Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE* 2011; **6**: e23610.

4. **Valdivia A, et al.** Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks – results for 2009–10. *Eurosurveillance* 2010; **15**: pii=19621.

5. **Hulth A, Rydevik G.** Web query-based surveillance in Sweden during the influenza A(H1N1)2009 pandemic, April 2009 to February 2010. *Eurosurveillance* 2011; **16**: pii=19856.

6. **Wilson N, et al.** Interpreting Google flu trends data for pandemic H1N1 influenza: the New Zealand experience. *Eurosurveillance* 2009; **14**: pii=19386.

7. **Kelly H, Grant K.** Interim analysis of pandemic influenza (H1N1) 2009 in Australia: surveillance trends, age of infection and effectiveness of seasonal vaccination. *Eurosurveillance* 2009; **14**: pii=19288.

8. **Malik MT, et al.** 'Google flu trends' and emergency department triage data predicted the 2009 pandemic H1N1 waves in Manitoba. *Canadian Journal of Public Health* 2011; **102**: 294–297.

9. **Thompson LH, Malik MT, Gumel A, Strome T, Mahmud SM.** Emergency department and 'Google flu trends' data as syndromic surveillance indicators for seasonal influenza. *Epidemiology and Infection* 2014; **142**: 2397–2405.

10. **The Flu Trends Team.** Google research blog. 20 August 2015. (http://googleresearch.blogspot.ca/2015/08/the-next-chapter-for-flu-trends.html). Accessed 29 September 2015.

11. **Google.** Google Flu Trends (http://www.google.org/flutrends/about). Accessed 6 October 2015.

12. **Google.** Google Flu Trends weekly influenza activity estimates for the world (http://www.google.org/flutrends/data.txt). Accessed 6 July 2015.

13. **Google.** Google Flu Trends – Canada (https://www.google.org/flutrends/ca/data.txt). Accessed 6 July 2015.

14. **WayBackMachine.** (http://archive.org/web/).

15. **Lazer D, et al.** Big data. The parable of Google Flu: traps in big data analysis. *Science* 2014; **343**: 1203–1205.

16. **The Official google.org blog.** (http://blog.google.org/). Accessed 6 July 2015.

17. **Google.** Frequently asked questions (http://www.google.org/flutrends/about/faq.html). Accessed 26 March 2015. [*Note: this website is no longer active, but can be accessed using the WayBackMachine.*].

18. **Public Health Agency of Canada.** Weekly FluWatch Reports Archive (http://www.phac-aspc.gc.ca/fluwatch/archive-eng.php). Accessed 26 March 2015.

19. **Public Health Agency of Canada.** Lessons Learned Review: Public Health Agency of Canada and Health Canada Response to the 2009 H1N1 Pandemic. 2010 (http://www.phac-aspc.gc.ca/about_apropos/evaluation/reports-rapports/2010-2011/h1n1/pdf/h1n1-eng.pdf).

20. **R Core Team.** R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2016.

21. **Hartley M.** Google Flu Trends to track spread of influenza in Canada in real time. *National Post*, 8 October 2009 (http://www.financialpost.com/related/topics/google+trends+track+spread+influenza+canada+real+time/2079477/story.html). Accessed 6 July 2015.

22. **Mohebbi M, Vanderkam D.** Google Flu Trends expands to 16 additional countries. Google official blog, 8 October 2009 (http://googleblog.blogspot.ca/2009/10/google-flu-trends-expands-to-16.html). Accessed 6 July 2015.

23. **Google.** Google Flu Trends weekly influenza activity estimates for the world (http://web.archive.org/web/20101231104329/http://www.google.org/flutrends/data.txt). [*Note: these are the 'revised' estimates.*].

24. **Google.** Google Flu Trends weekly influenza activity estimates for the world (http://web.archive.org/web/20100915104329/http://www.google.org/flutrends/data.txt). [*Note: these are the 'original' estimates.*].

25. **The Official google.org blog.** Comparing flu around the world, 12 November 2010 (blog.google.org/2010/11/comparing-flu-around-world.html).

26. **Public Health Agency of Canada.** FluWatch 19 April 2009 to 25 April 2009 (week 16) (http://web.archive.org/web/20090817172209/http://www.phac-aspc.gc.ca/fluwatch/08-09/w16_09/index-eng.php). Accessed 28 August 2016.

27. **Newfoundland and Labrador Department of Health and Community Services.** Influenza Weekly Report 8 March–14 March, week 10, 2014–2015 (http://www.health.gov.nl.ca/health/publichealth/cdc/flu/NL_Influenza_Report_Week%2010.pdf). Accessed 19 November 2015.

28. **Manitoba Health, Healthy Living, and Seniors.** Influenza Surveillance Weekly Report 2014/2015 season – week 8, 22 February–28 February 2015 (http://www.gov.mb.ca/health/publichealth/surveillance/influenza/docs/150228.pdf). Accessed 19 November 2015.

L. J. MARTIN
University of Alberta, Edmonton, Alberta, Canada

Address for correspondence:
(Email: leah.martin@ualberta.ca; leahjmartin1@gmail.com)