Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season

Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season

David Lazer,[1,2]* Ryan Kennedy,[1,2,3] Gary King,[2] Alessandro Vespignani[1]

[1]Northeastern University, Boston, MA 02115, USA. [2]Harvard University, Cambridge, MA 02138, USA. [3]University of Houston, Houston, TX 77204, USA.
*Corresponding author. E-mail: d.lazer@neu.edu.

Last year was difficult for Google Flu Trends (GFT). In early 2013, *Nature* reported that GFT was estimating more than double the percentage of doctor visits for influenza-like-illness than the Centers for Disease Control and Prevention's (CDC) sentinel reports during the 2012-2013 flu season (*1*). Given that GFT was designed to forecast upcoming CDC reports, this was a problematic finding. In March 2014, our report in *Science* found that the overestimation problem in GFT was also present in the 2011-2012 flu season (*2*). The report also found strong evidence of autocorrelation and seasonality in the GFT errors, and presented evidence that the issues were likely, at least in part, due to modifications made by Google's search algorithm and the decision by GFT engineers not to use previous CDC reports or seasonality estimates in their models – what the article labeled "algorithm dynamics" and "big data hubris" respectively. Moreover, the report and the supporting online materials detailed how difficult/impossible it is to replicate the GFT results, undermining independent efforts to explore the source of GFT errors and formulate improvements.

To address the accuracy problems from the 2012-2013 flu season, GFT engineers announced modifications to the algorithm at the annual conference of the International Society of Neglected Tropical Diseases (*3*). These modifications relied on the assumption that increased media coverage of the flu during the 2012-2013 season was the cause of the error – an assumption shared by most all of the media coverage of the problem (*1, 4, 5*). Two changes were made: (1) dampening anomalous media spikes and (2) using ElasticNet, rather than regression, for estimation.

GFT still stands as a triumph of big data engineering. This is why it is so critical that continued monitoring and re-evaluation of the results take place, not just within the GFT team but also within the larger academic community.

So have these changes corrected the problem? While it is impossible to say for sure based on one subsequent season, the evidence so far does not look promising. First, the problems identified with replication in GFT appear to, if anything, have gotten worse. Second, the evidence that the problems in 2012-2013 were due to media coverage is tenuous. While GFT engineers have shown that there was a spike in coverage during the 2012-2013 season, it seems unlikely that this spike was larger than during the 2005-2006 A/H5N1 ("bird flu") outbreak and the 2009 A/H1N1 ("swine flu") pandemic. Moreover, it does not explain why the proportional errors were so large in the 2011-2012 season. Finally, while the changes made have dampened the propensity for overestimation by GFT, they have not eliminated the autocorrelation and seasonality problems in the data.

Making the Black Box Still Darker?

One of our main concerns about GFT is the degree to which the estimates are a product of a highly non-transparent process. Science is a cumulative enterprise, and progress requires the ability for the community to continually assess the work on which they are building (*6*, *7*). GFT has not been very forthcoming with this information in the past, going so far as to release misleading example search terms in previous publications (*2*, *3*, *8*).

These transparency problems have, if anything, become worse. While the data on the intensity of media coverage of flu outbreaks does not involve privacy concerns, GFT has not released this data nor have they provided an explanation of how the information was collected and utilized. This information is critically important for future uses of GFT. Scholars and practitioners in public health will need to be aware of where the information on media coverage comes from and have at least a general idea of how

it is applied in order to understand how to interpret GFT estimates the next time there is a season with both high flu prevalence and high media coverage.

Within the new data released by GFT, there are also issues. GFT released backdated data based on their new algorithm, but this data does not seem to coincide with the data released on their main page (*9*). For example, on October 6, 2013, the main Google Flu Trends download page reports that 1.405% of doctor visits were for ILI – a number that is reportedly based on the 2013 algorithm and which is relatively close to the CDC's estimate of 1.2738%. The backdated data, which is reportedly done with the same algorithm reports a prevalence of only 0.328%. Backdated information for the 2009 algorithm also does not line up with what is reported on GFT's main page. On September 20, 2009, for example, the posting on the main page (based on the 2009 updated algorithm) reports an estimated prevalence of 4.254%, again very close to the CDC's estimate of 4.2158%. The backdated data file shows an estimate of 1.458% (*10*).

After much digging, we were able to identify what was wrong with the 2009 algorithm's back data. The file labeled as the "2009 Model Update applied to prior years" is actually the original 2008 algorithm's results. We were only able to confirm this because we located an older version of the GFT website that had been crawled by the Internet Archive (aka. "The Wayback Machine") (*11*). We have no explanation for the inconsistency with the 2013 model update file. We have archived these pages and can make them available on request.

When we attempted to contact Google.org about this issue, we found it difficult to get a response, or to identify who to contact. The feedback form provided on the GFT website leads to a page saying that they will "not be able to reply to any feedback or requests for support" (Fig 1). The academic paper outlining the 2013 update provides no contact author, only stating, "For more information, please contact: google.org" (*3*). We attempted to do so, and also attempted to find e-mail addresses for any of the authors listed in the paper, and were unsuccessful in all these attempts. We appear not to be the only people to have encountered this problem. A Google search for "contacting google.org" shows a post to Google Groups asking how to contact google.org that has no replies and is closed "due to inactivity" (*12*). There is no readily identifiable contact source for information on how to reconcile these differences in data that is reportedly generated by the same algorithm and data.

Barking Up the Wrong Tree?

To explain the problems in GFT performance in the 2012-2013 flu season, the GFT engineers posited a straightforward explanation: "We hypothesized that concerned people were reacting to heightened media coverage, which in turn created unexpected spikes in the query volume. This assumption led to a deep investigation into the algorithm that looked for ways to insulate the model from this type of media influence" (*3*).

But how likely is this explanation? GFT presents the chart in Fig 2 as evidence. The figure shows the relationship between GFT's absolute error rate and news coverage of the flu – although we are never told from where the news coverage measure comes. Noticeably, it is not possible to detect the media panics associated with the 2005-2006 A/H5N1 ("bird flu") outbreak and the 2009 A/H1N1 ("swine flu") pandemic in this chart. The chart appears to plot news volume only beginning in late 2010, even though the chart claims to start the comparison in 2004.

There are other problems with this explanation. When evaluating their misses, GFT engineers use the raw error of the GFT estimate (GFT-CDC). This, however, is not a good way to visualize the error because it tends to over-emphasize errors during years with higher flu prevalence. In the first week of 2012, the GFT estimate of ILI prevalence was also about 64.7% higher than the CDC estimate. Because the CDC estimate at that time was much lower, the absolute error was only about 19.4% the size of the error in the first week of 2013. Fig 3 compares the absolute and the proportional ((GFT-CDC)/CDC) error over these two flu seasons. When this is done, it becomes clear that problems were present well before the heightened media attention reported in Fig 2.

How did the updated GFT algorithm perform in the 2013-2014 season? It certainly succeeded in lowering the persistently high GFT estimates of ILI. There are six weeks of overlap between our initial download of GFT data (in September) and the updated version (downloaded in March). The correlation between the two is still high (r=0.962), but the new estimates are, on average, about 12.4% lower. This is also reflected in the overall propensity for GFT to estimate higher than the CDC. Whereas in the original data, GFT estimated high 100 out of 108 weeks from August 21, 2011 to September 1, 2013 (about 92.6% of the time), the updated estimates are higher than the CDC estimates 23 out of 31 weeks (about 74.2% of the time).

Fig 4 shows the comparison between the CDC and GFT estimates this past season. The errors are indeed lower, but GFT still estimated a peak that was about 30% higher than the CDC's estimate. The new algorithm has clearly succeeded in dampening the GFT estimates, but it remains an open question whether it has done so by better capturing the relationship between search and ILI prevalence, or if the changes have just generally dampened the signal.

Still Information Available?

More worrying is that the patterns of autocorrelation and seasonality in the errors are still quite clear even from a casual glance at the data. More systematic checks for autocorrelation confirm this intuition ($p < 0.05$) (Fig 5). While checks for seasonality are not feasible for only one season, it is notable that the peak error continues to occur around the peak of the flu season.

What this means is that there is still substantial information that is not being utilized by GFT. In the thirty weeks since the new GFT was launched, the mean absolute error (MAE) was 0.211. It has started to perform better than the simplest model, which uses only a single two-week lag of the CDC's data to predict future reports (*13*), but not significantly so ($p > 0.25$, sign test, see *14*). Similarly, for a model including only CDC lags and seasonality variables (*15*), the MAE is 0.233 – higher, but not significantly so ($p > 0.50$). However, as was documented in previous studies, adding the lagged difference between the GFT and CDC data (*16*) reduces the MAE substantially, to 0.085 – also not statistically significant because of the low number of cases, but nearly so ($p < 0.20$).

More leverage can be gained by looking at GFT's estimates for the 10 Health and Human Services (HHS) regions used to help develop the model (*1*). Fig 6 shows a test for autocorrelation within the GFT errors for each of these regions. This chart shows significant levels of autocorrelation ($p < 0.05$) for every region except Region 8. Again, with only 30 time points since the new algorithm's launch, the detection of significant autocorrelation at such an early stage is not encouraging.

Moreover, we can compare the differences across these regions to expand the size of our dataset (150, rather than 30 observations) and get a better handle on whether the differences observed on the

national level are significant. Fig 7 compares the MAE for GFT, the single-lag CDC model, the more complex lagged CDC model with seasonality, and the model which incorporates CDC, GFT and seasonality data. There are substantial differences in performance between regions. GFT generally performs better than the models which only use CDC and/or seasonality data, but these differences are not that large and they are not statistically significant ($p > 0.35$ for both). On the other hand, the model that utilizes all of the data sources performs about as good or (much) better in every region except Region 8. Across the regions, this improvement is statistically significant ($p < 0.001$).

What these result suggest is that GFT is still not performing significantly better than simple CDC lagged models. It also indicates that GFT is leaving substantial information out of its estimates that could result in improved estimates on both the regional and national level. It may still be too early to state these results with complete certainty. This being said, we should note that the results show a substantial improvement from explicitly modeling the systematic error pattern in GFT – and this may avert problems that arise with GFT in the future.

What Does the Future Hold?

While it is still far too early to say whether the patches applied by the GFT engineers will hold in future flu seasons, this report suggests reasons for skepticism. The argument that high media coverage was the primary cause of GFT's 2012-2013 problems is tenuous. Worse, by dampening the signals from searches during intense media coverage periods, GFT might encounter problems during periods of both high media coverage and high flu prevalence. Similarly, GFT is still ignoring data that could help it avoid future problems.

References And Notes

1. D. Butler, *Nature* **494**, 155 (2013).
2. D. Lazer et al., *Science* **Forthcoming**, na (2014).
3. P. Copeland *et al*., *Int. Soc. Negl. Trop. Dis.* **2013**, 3 (2013).
4. T. Claburn, *InformationWeek*, http://www.informationweek.com/how-google-flu-trends-blew-it/d/d-id/1112081 (2013).
5. N. Statt, *CNET*, http://news.cnet.com/8301-11386_3-57614392-76/flu-predictions-get-more-accurate/ (2013).
6. G. King. Replication, Replication. *PS* **28**, 443-499 (1995).
7. P. Voosen. Researchers struggle to secure data in an insecure age. Chronicle of Higher Education, http://chronicle.com/article/Researchers-Struggle-to-Secure/141591/ (2013).
8. S. Cook et al. Assessing Google Flu Trends performance in the United States during the 2009 Influenza Virus A (H1N1) pandemic. *PLoS One* **6**: e23610. doi:10.1371/journal.pone.0023610 (2011).
9. http://www.google.org/flutrends/historic/us-historic-v2.txt
10. http://www.google.org/flutrends/historic/us-historic-v1.txt
11. http://web.archive.org/web/20090303211715/http://www.google.org/about/flutrends/download.html
12. https://productforums.google.com/forum/#!topic/apps/SmPE9JEkCz8
13. $flu_t = \alpha + \beta_1 flu_{t-2}$, where *flu* is the CDC sentinel estimate of ILI prevalence.
14. F.X. Diebold and R.S. Mariano. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* **13**, 253-265 (1995).
15. $flu_t = \alpha + \beta_1 flu_{t-2} + \beta_2 flu_{t-3} + \beta_3 flu_{t-4} + \sum_{i=1}^{52} \gamma_i \, week_{it}$, where *flu* is the CDC sentinel estimate of ILI prevalence, *week* is the week of the year, and *β* and *γ* are regression parameters.
9. $flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-2} + \beta_3 (gflu_{t-2} - flu_{t-2}) + \beta_4 (gflu_{t-3} - flu_{t-3}) + \sum_{i=1}^{52} \gamma_i week_{it}$, where *flu* is the CDC sentinel estimate of ILI prevalence, *gflu* is the GFT estimate of ILI prevalence, *week* is the week of the year, and *β* and *γ* are regression parameters.

Acknowledgements: This research was funded, in part, by NSF grant no. 1125095 and, in part, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract D12PC00285. We also gratefully acknowledge the help and support provided by HRL Laboratories, LLC. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, IARPA, DoI/NBE, HRL, or the U.S. government.

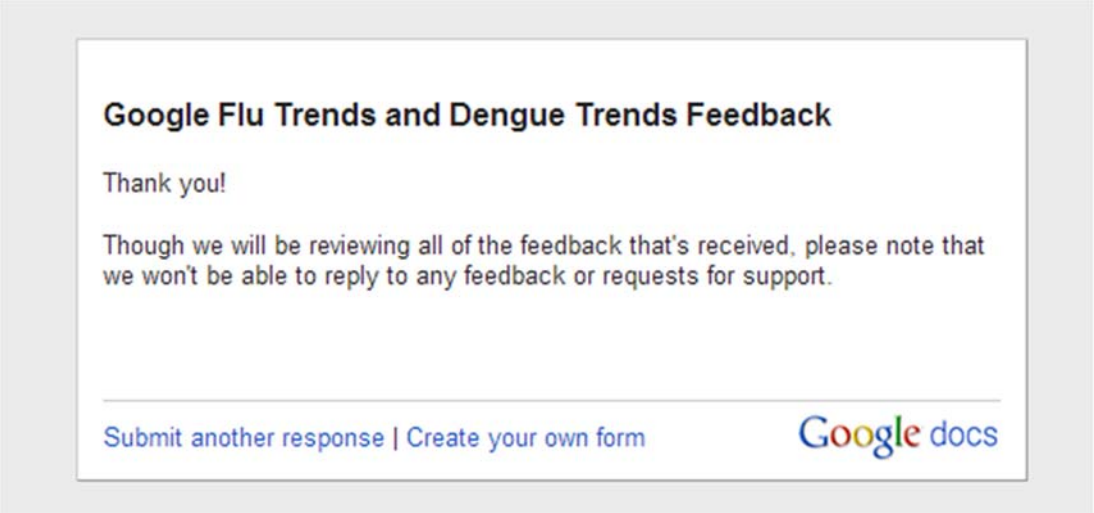


Fig 1. Result of Submitting Feedback Form With Request for Clarification on Google Flu Trends' FAQ Page. This was the message we received after submitting a request for clarification on the seeming inconsistency between the data available for download on GFT's main page (http://www.google.org/flutrends/) versus the historic time series posted on their "How does this work?" page (http://www.google.org/flutrends/about/how.html). The form can be found on their FAQ page under the title "How can I sent feedback about Google Flu Trends?" (http://www.google.org/flutrends/about/faq.html). Attempts to find a general contact through google.org, as recommended by [*3*] proved unsuccessful.

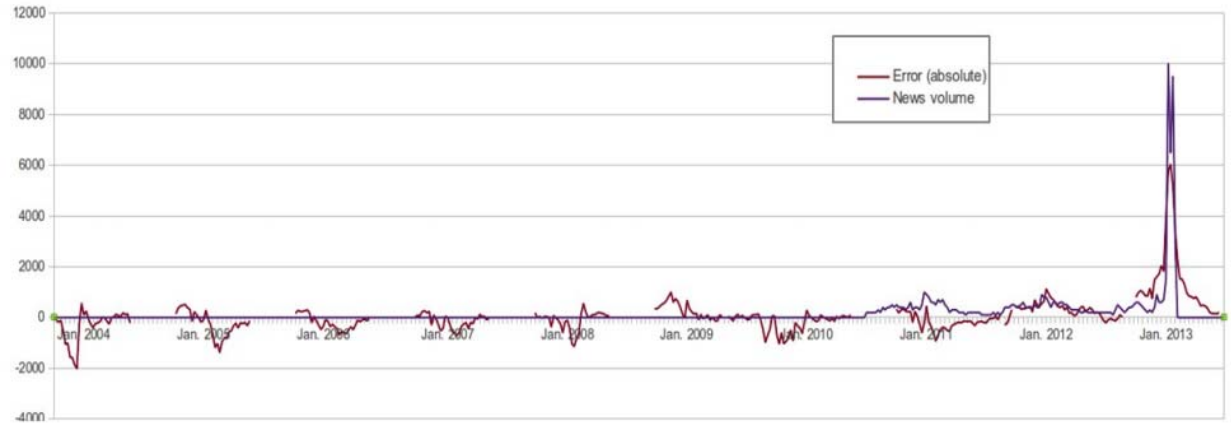


Fig 2. GFT plot of media volume and prediction error rate, 2004-2013. Chart drawn directly from GFT update report (*3*). While the chart purports to show both GFT error and news coverage from 2004 to 2013. However, it appears that news volume is only measured after about June 2010. This means that the media panics around the 2005-2006 A/H5N1 ("bird flu") outbreak and the 2009 A/H1N1 ("swine flu") pandemic are not included – these are also periods where GFT's 2009 algorithm performed relatively well.

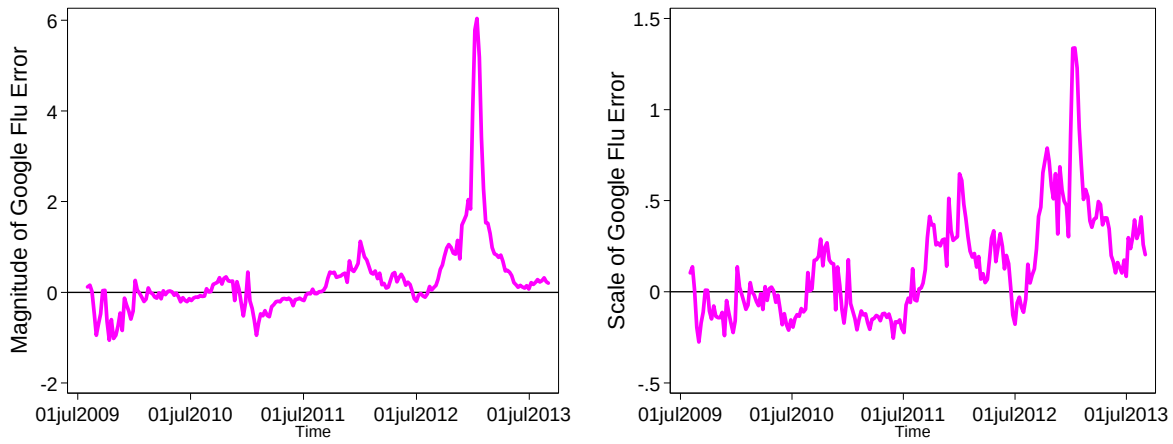


Fig 3. Comparison of absolute and proportional error in the 2010-2013. Absolute error of GFT versus CDC (GFT-CDC) (**Left**) and proportional error of GFT [(GFT-CDC)/CDC] (**Right**). Scaling the errors paints a different picture of GFT errors than looking at absolute error. Absolute error will look larger when the baseline level of flu is lower. This is why most of the focus has been on 2012-2013 as an aberration. Scaled error reveals that GFT also predicted flu prevalence that is about 65% higher than the CDC estimate in 2011-2012, and missed high in 100 out of 108 weeks from August 21, 2011 to September 1, 2013. All of this suggests that GFT's problems started earlier than is usually thought and might not correlate as highly with spikes in media coverage.

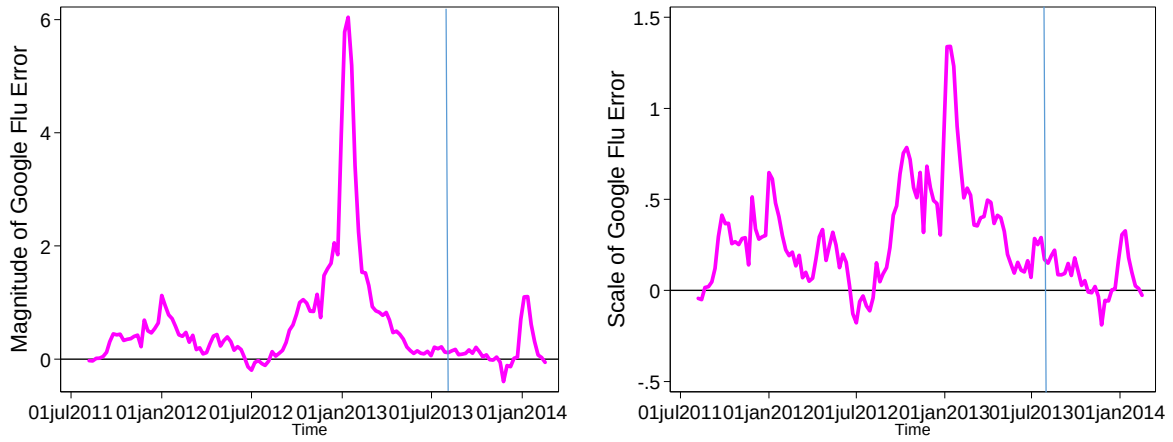


Fig 4. GFT vs. CDC estimates after latest GFT update. Absolute (**Left**) and proportional (**Right**) error of GFT from 2011 to 2014. Observations to the right of the blue line are after GFT started its new algorithm. While the update has dampened the size of GFT estimates (by about 12% for those observations in which we have overlap between the old and new model), GFT is still estimating high almost 75% of the time. It also still estimated about 30% higher than the CDC in the 2013-2014 flu season.

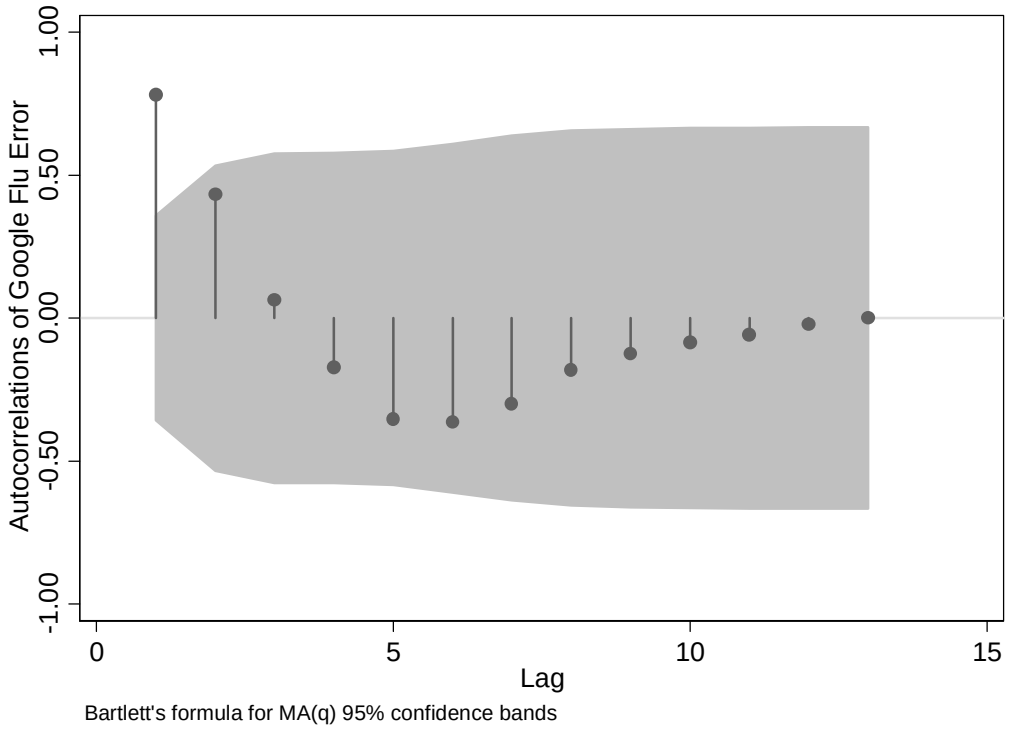


Fig 5. Autocorrelation plot for GFT post-update errors, post-September 22, 2013. A correlogram (graph of autocorrelations) of GFT errors since the latest update. 95% confidence intervals (shaded region) based on Bartlett's formula for MA(q) processes. Even for this short time period (30 weeks) the test for an AR(1) process rejects the null hypothesis of no autocorrelation at the 95% level.

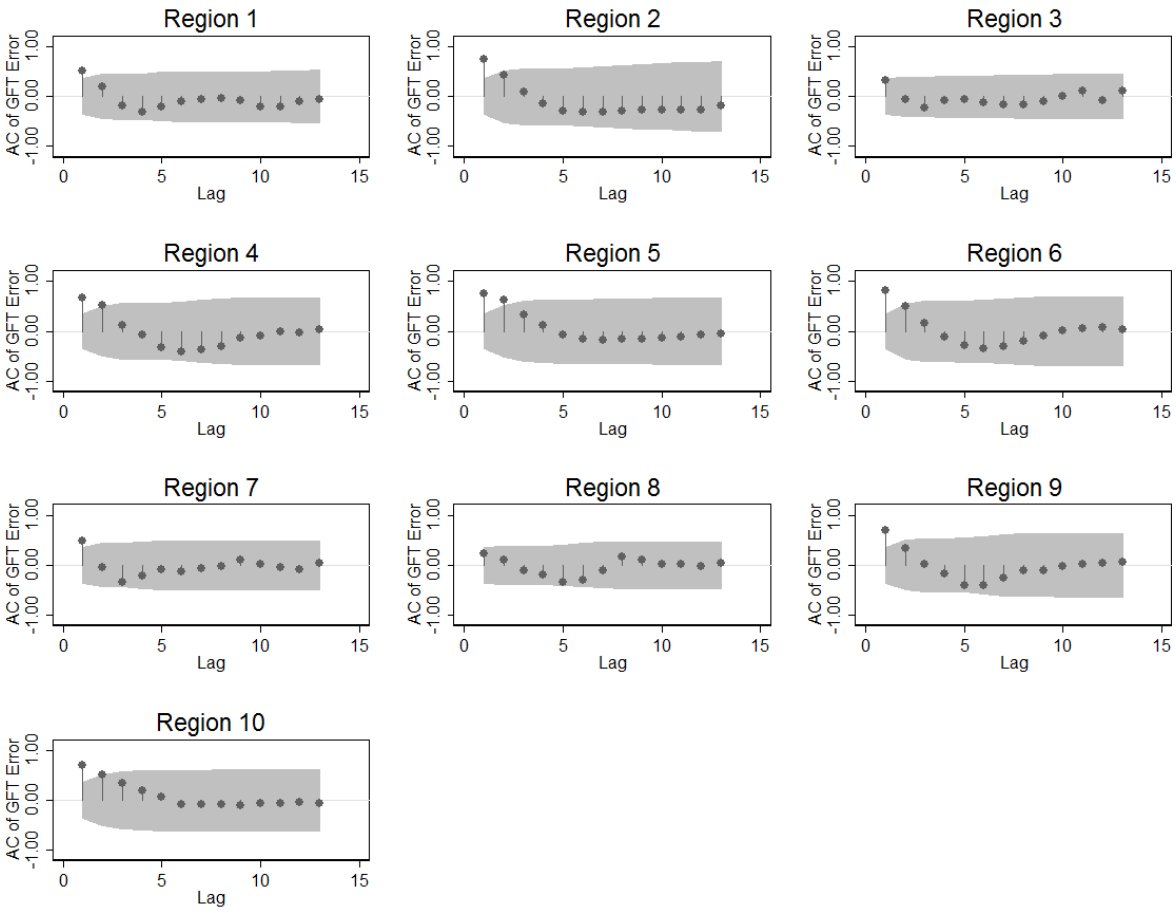
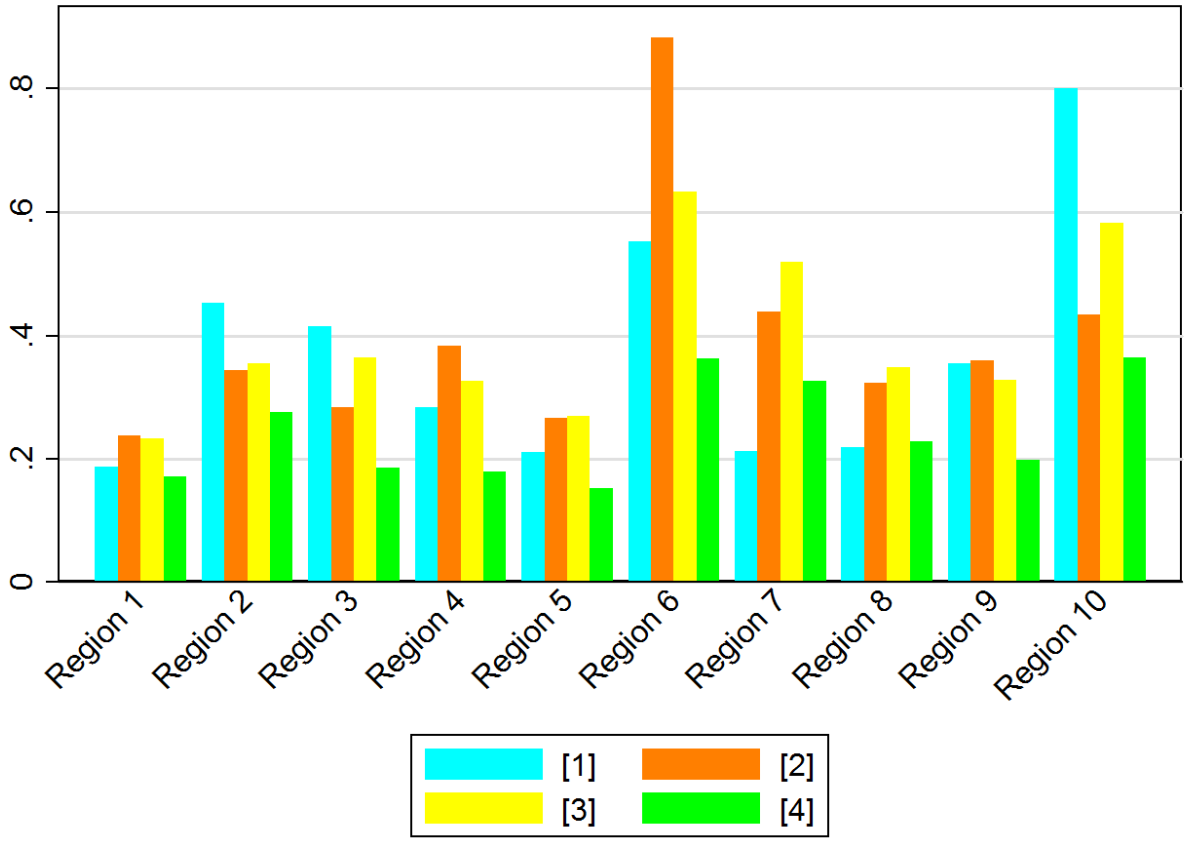


Fig 6. Autocorrelation Plot of Regional GFT post-update errors, post-September 22, 2013.
Correlograms (graph of autocorrelations - AC) of GFT errors in the 10 HHS regions since the latest update. 95% confidence intervals (shaded region) based on Bartlett's formula for MA(q) processes. Even for this short time period (30 weeks) the test for an AR(1) process rejects the null hypothesis of no autocorrelation at the 95% level for 9 or the 10 regions.



[1] $flu_t = gflu_t$
[2] $flu_t = \alpha + \beta_1 flu_{t-2}$
[3] $flu_t = \alpha + \beta_1 flu_{t-2} + \beta_2 flu_{t-3} + \beta_3 flu_{t-4} + \sum_{i=1}^{52} \gamma_i \, week_{it}$
[4] $flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-2} + \beta_3 (gflu_{t-2} - flu_{t-2}) + \beta_4 (gflu_{t-3} - flu_{t-3}) + \sum_{i=1}^{52} \gamma_i week_{it}$

Fig 7. Regional Error for 2013-2014 Season. Comparison of mean absolute error (MAE) for GFT versus three alternative models after GFT's most recent update (September 22, 2013). Note that performance for the models varies considerably across regions, but the differences between GFT's estimates and estimates using only CDC data or a combination between CDC data and seasonality variables are only substantial in a couple of regions. Contrariwise, the MAE for the model combining GFT's estimates with CDC data and seasonality variables produces much lower error in almost every region except 7 and 8, but here the differences are relatively small.