How Not to Forecast the Flu

Prithwish Chakraborty[1,2*], Bryan Lewis[3], Stephen Eubank[3], John S. Brownstein[4,5], Madhav Marathe[2,3], and Naren Ramakrishnan[1,2],

1 Discovery Analytics Center, Virginia Tech, VA, USA
2 Dept. of Computer Science, Virginia Tech, VA, USA
3 Network Dynamics and Simulation Science Laboratory, Biocomplexity Institute, Virginia Tech, VA, USA
4 Children's Hospital Informatics Program, Boston Children's Hospital, MA, USA
5 Dept. of Pediatrics, Harvard Medical School, MA, USA

* prithwi@vt.edu

Summary

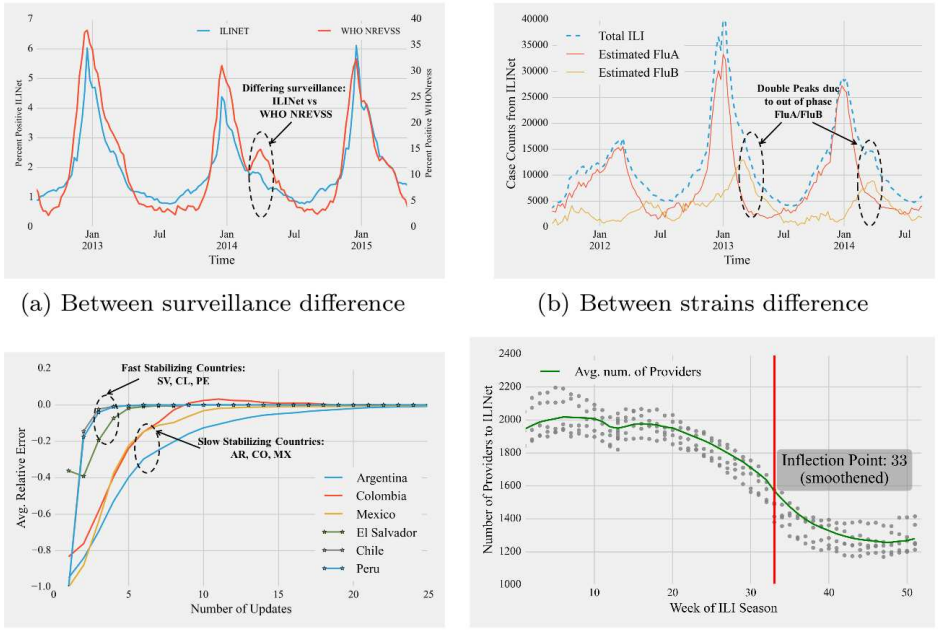
Accurate and timely influenza (flu) forecasting has gained significant traction in recent times. If done well, such forecasting can aid in deploying effective public health measures. Unlike other statistical or machine learning problems, however, flu forecasting brings unique challenges and considerations stemming from the nature of the surveillance apparatus and the end-utility of forecasts. This article presents a set of considerations for flu modelers that must be addressed for effective and useful forecasting.

Introduction

As we prepare to embark upon a new flu season, we will hear the usual cautionary notes about vaccinations, the preparedness of our health systems, and the specific strains that are relevant for these seasons. Recent competitions, organized by agencies like the CDC and IARPA, have spurred interest in flu forecasting across academia and industry. While the CDC competition aimed to forecast flu seasonal characteristics in the US, the IARPA Open Source Indicators (OSI) forecasting tournament was focused on disease forecasting (flu and rare diseases) in countries of Latin America. Our team was declared the winner in the IARPA OSI competition and our goal here is to communicate our lessons learned about what goes into a successful forecasting engine while ensuring its relevance to public health policy and, consider some common assumptions that can be violated and require careful attention. Broadly, such considerations can be categorized with respect to 'Surveillance Characteristics' (see Fig 1) and 'Forecasting Practices' (see Fig 2). We discuss these considerations and present our recommendations about the pitfalls to avoid, as below.

Surveillance Characteristics

Flu surveillance networks exhibits several unique traits which can further be modulated by regions/time period of interest. An effective flu forecasting system needs to pay careful attention to such characteristics. We discuss some of the more important, but often overlooked, aspects of the surveillance characteristics in the following sections.



(a) Between surveillance difference

(b) Between strains difference

(c) Surveillance Instability

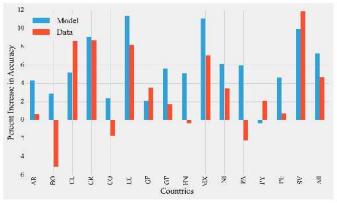
(d) Surveillance drop-off

Fig 1. Different Characteristics of Surveillance systems. (a) Phase differences in reported ILI percentages between two surveillance systems(ILINet vs WHO NREVSS). (b) Phase differences between different ILI substrains (Flu A vs Flu B) leading to double peaked overall ILI curve. (c) Surveillance reports unstable many weeks after first report. While countries like Chile stabilizes quickly (within 5 weeks), other countries like Argentina stabilizes after many weeks (≥ 10). (d) Surveillance drop-off towards the end of the season - scatter plot of Number of providers reporting to CDC ILINet as a function of ILI season week. Green Line shows the smoothened average while the red vertical line shows the smoothened inflection point of surveillance coverage. (Smoothing interval = 4)

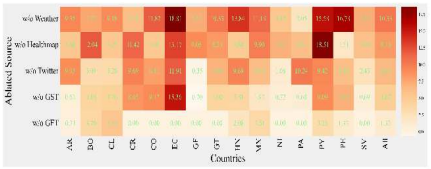
Surveillance networks do not measure the same quantity

Influenza-like Illnesses (ILI), tracked by many agencies such as CDC, PAHO, and WHO [1–3], is a category designed to capture severe respiratory disease, like influenza (flu), but also includes many other less severe respiratory illness due to their similar presentation. Surveillance methods often vary between agencies. Even for a single agency, there may be different networks (such as outpatient based and lab sample based) tracking ILI/Flu. While outpatient reporting networks such as ILINet aim to measure exact case counts for the regions under consideration, lab surveillance networks such as WHO NREVSS (used by PAHO) seek to confirm and identify the specific strain. In the absence of a clinic based surveillance system, lab-based systems can provide estimates at per "X" population level; however making an estimate of actual influenza flu cases from these systems is challenging [1]. Furthermore, surveillance reports are often non-representative of actual ILI incidence (see. "epidemic data pyramid") and can often suffer from variations such as holiday periods where behavior of people visiting hospitals changes from other weeks (see. "Christmas effect" in supplementary material). Additionally, the number of cases represented by a single positive lab sample in a densely populated country with wide lab network coverage necessarily differs from one in a more resource constrained lab surveillance system. For

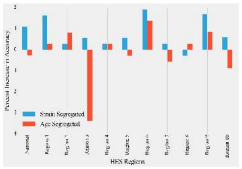
(a) Single source prediction accuracy



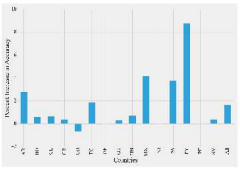
(b) Multiple source prediction accuracy



(c) Percentage Accuracy reduction due to source ablation



(d) Segregation: Prediction accuracy



(e) Instability Correction

Fig 2. Forecast accuracy under various conditions. (a) *Single Source:* Forecast accuracy for each individual source (Weather, Healthmap, Twitter, Google Flu Trends and Google Search Trends). No particular source is the best for all countries. (2) *Multiple Sources:* Percent increase in forecast accuracies while combining multiple sources at Model level and at Data level over best single source forecasts. Model level gives better overall performance. (c) *Ablation test:* Percent reduction in forecast accuracies while removing one source at a time from the Model fusion. Removing a source can lead to better performance for some countries. (d) *Segregation Test:* Percent increase in forecast accuracies for US ILI data considering segregation by age and by subtype over unsegregated forecasts. Segregated methods show better accuracy. (e) *Instability Correction:* Percent increase in forecast accuracies for different countries after correction over uncorrected forecasts. Significant improvement can be seen for countries like Argentina and Paraguay.

example, in the US, the CDC reports percentage positive ILI for ILINet (an outpatient reporting network) and for WHO NREVSS (based on lab samples). While both networks are designed for different purposes and provide valuable information for forecasting, they exhibit significant phase differences (see Fig 1a). and, it is essential for forecasters to distinguish between surveillance networks and their intricacies. This difference is especially important when conducting cross-correlation studies between different countries (e.g., geographically nearby tropical countries may have similar underlying ILI activity) or even between two different networks for the same country.

Flu is not ILI. ILI is plural. Geographical diversity must be modeled

Many surveillance systems for influenza report on Influenza like Illnesses (ILI) [1] which is based on symptoms: fever (temperature of 100°F or greater) and cough and/or sore throat. This general classification is easier to gather than more expensive and labor-intensive virological tests; however, this definition means that multiple influenza strains and respiratory infections are aggregated together. Such aggregation poses difficulties in modeling because different strains of the flu exhibit different seasonal characteristics (see Fig 1b) where Flu A shows significantly different phase than Flu B). Consequently, a single forecasting model is unlikely to capture the behavior of the overall epicurve. Our recommendation is to not treat ILI as an atomic

illness and instead predict for sub-strains if data at required resolution(s) is available. Our experiments suggest that such breakdown by strains can produce better quality predictions by accounting for phase shifts between them (see Fig 2d).

Similarly, building models at the national level without considering differences in geography is likely to lead to erroneous results as ILI characteristics are again shifted and scaled differently across regions. In a geographically diverse country such as the United States, ILI seasonal curves are consistently out of phase between different HHS regions as well as the national curve. Further, a single region (e.g., HHS region 10 which subsumes states such as Alaska and Washington) can be composed of non-contiguous land masses which may make any assumption about uniformity in ILI profiles inaccurate.

Surveillance data is not stable

Epidemiologists who work with surveillance data know that, while useful, they are often delayed and can be candidates for revision/updating for several weeks after initial publication. The lag between initial publication and final revision can be as small as 2 weeks (e.g., for CDC ILINet data) or can wildly fluctuate (for example, PAHO reports for some Latin American countries such as Argentina, Colombia and Mexico can take more than 10 weeks to settle. On the other hand, PAHO reports stabilize within 5 weeks for countries such as Chile, Costa Rica and Peru (see Fig 1c). The reason for such discrepancies has to do with the maturity of the surveillance apparatus and the level of coordination underlying public health reporting. Most works on forecasting do not account for such instability. In essence, we are forecasting a moving target and recent research [4] suggests that modeling such uncertainty directly in a forecasting model can significantly improve performance. In particular, a stabilization or auto-correction term can be inferred by conducting a regression for the final estimate as a function of the intermediate values (and their times of publication)(see Fig 2e). One could devise more intricate algorithms that take into account cross-country correlations and country-specific societal factors to correct for imperfections in surveillance.

Surveillance data collection practices are not uniform

Even within a surveillance framework, there are systematic deviations in coverage as well as differential lags in updates. Surveillance reporting has been known to taper off or stop altogether during the post-peak part of the season. For example, as is evident from Fig 1d, the number of providers who reported to US CDC ILINet surveillance tapers off towards the end of the ILI season week (for US, calendar week 40 corresponds to first ILI season week [1]). Specifically, the inflection point of the average curve occurs at week 33. Such effects can possibly be attributed to resource re-allocation due to reduced interest in post-peak activities. This necessitates modelers to make a distinction between expected ILI/flu curves and the observed data for different parts of the season. Some efforts have been made to account for such systematic deviations by either correcting their estimates to match the surveillance reports [5] or by explicitly modeling surveillance errors [6]. More explicit data about the detailed structure of the surveillance system would benefit these corrections.

Forecasting Practices

There is no community agreement on measure(s) of performance – forecasting reason

Measuring forecasting skill is dependent on the actual use of the forecasts, which varies widely. Not surprisingly, there is no accepted measure of forecasting performance. We conducted a survey [7] where we identified 7 different metrics for around 10 different quantities each evaluating a different facet of flu. Moreover, evaluations often involve multiple criteria, can include subjective components, and present trade-offs where the balance of preferences is not well articulated. A vanilla mean-squared error criterion will lead to a model with a tendency to under-predict the peaks when trained uniformly over complete seasons. However, if agencies are more interested in peak characteristics such a measure can be modified to penalize deviations around peak more severely. Thus understanding the requirements of health agencies is essential to generate meaningful forecasts. Additionally, quantifying forecasting uncertainty and presenting the same [8] in a meaningful manner is of prime importance to facilitate actionable strategies from health agencies.

The best case count predictor might not capture seasonal characteristics

Forecasting tournaments provide much needed impetus to improving the state-of-the-art of the field but viewing ILI prediction as machine learning or regression problem misses the big picture. In particular, the best algorithm according to a sum-of-squared-error or classical goodness-of-fit measure might completely miss the most important seasonal characteristics such as the start epiweek of the season, peak epiweek, and peak value - indicators likely to be more important for public health and infection control purposes.

Most published literature focuses on retrospective evaluations of forecasting methods

To the best of our knowledge, the CDC and IARPA OSI competitions are the only two competitions that truly involve forecasting into the future. Participants were required to submit time-stamped forecasts that must arrive before the event date (and thus certainly before the surveillance data becomes available). In contrast, almost all published literature focuses on retrospective analysis on flu seasons past [9]. The benefit of hindsight affords the ability to tune models indefinitely, and out-of-sample testing is no substitute for forecasting. Dangers include not just overfitting but also vulnerability to model drift over time.

Many forecasting projects are not predicting into the future; they are really nowcasting, or near-casting

Due to the delays and updates inherent in surveillance data, a "forecasting" project predicting one or two weeks into the future is in essence a "nowcasting" [10] system. Understanding this aspect is crucial as it could lead to intelligent strategies of handling surrogate information [11] such as weather data and social network sensors. Since surrogate sources are typically real-time, when performing nowcasting one can ideally complement knowledge of last known surveillance data with the current surrogates and hence find a direct estimate of the surveillance data. On the other hand a prediction engine interested in forecasting the peak of the flu season (say 7

weeks in advance), does not have the same liberty. Some researchers have used data assimilation techniques [10, 12] to incorporate knowledge about weather data to provide ILI forecasts. Intelligent assimilation techniques can be developed to incorporate weather forecasts (which are also available) and reduce the error bounds while predicting for more than, say, four weeks in advance.

More Data is not always Better Data

There is now a wealth of syndromic surveillance and physical indicators available for forecasting the flu such as weather, news reports, web search query volumes (GFT, GST), Twitter chatter [5] and Wikipedia logs [13, 14]. Pitfalls in big data analysis without understanding the underlying system biases have been well-documented (e.g., for GFT, see [15]). We evaluated diverse data sources for flu forecasting in 15 Latin American countries (Fig 2a) and have considered both data-level fusion and model-level fusion. The former concatenates all data sources and renders them in a common denominator format before modeling. Model-level fusion builds separate models with each source and combines their forecasts leveraging selective superiorities. As Fig 2b shows, model-level fusion performs better than data-level fusion. "Ablation tests" (see Fig 2c) show that physical indicators are consistently the more important predictor than social indicators. Inclusion of some data sources can actually lead to reduced performance for specific countries (e.g., Google Search Trends for French Guiana).

Conclusion

Infectious disease forecasting is a rapidly emerging field. The challenges reported here pertain to how data from surveillance systems can be modeled as well as how forecasting standards need to mature. While we have focused on flu forecasting here, these challenges translate to other diseases as well. We advocate increased efforts in fostering a community of forecasters and converging on a common understanding of shared concerns and approaches forward. We also argue for easier availability of surveillance data and surrogate sources in raw forms, at all possible spatial and temporal granularities. Skilled forecasting research can significantly aid in the translation of research results into the hands of public policy professionals and decision makers.

Supporting Information

S1 Appendix. Datasets and associated code snippets for this article is archived at the publicly accessible github page `http://prithwi.github.io/how_not2_flu`.

Acknowledgments

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337and by the Defense Threat Reduction agency (DTRA) via the CNIMS Contract HDTRA1-11-D-0016-000. The US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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 IARPA, DoI/NBC, DTRA, or the US Government.

References

1. CDC. Influenza (flu);. www.cdc.gov/flu/index.htm.
2. PAHO. Influenza and other Respiratory Viruses;. http://ais.paho.org/phip/viz/ed_flu.asp.
3. WHO. Surveillance and Monitoring;. http://www.who.int/influenza/surveillance_monitoring/en/.
4. Fischhoff B, Davis AL. Communicating Scientific Uncertainty. Proceedings of the National Academy of Sciences. 2014;111(Supplement 4):13664–13671.
5. Chakraborty P, Khadivi P, Lewis B, Mahendiran A, Chen J, Butler P, et al. Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions. In: Proceedings of SDM '14. SIAM; 2014. p. 262–270.
6. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-Time Influenza Forecasts during the 2012–2013 Season. Nature communications. 2013;4.
7. Matsubara Y, Sakurai Y, van Panhuis WG, Faloutsos C. FUNNEL: Automatic Mining of Spatially Coevolving Epidemics. In: Proceedings of KDD'14. ACM; 2014. p. 105–114.
8. Tabataba FS, Chakraborty P, Ramakrishnan N, Marathe MV, Chen J, Lewis BL. Standard Measures and Quality Metrics for Evaluating the Performance of Forecasting Methods: Special Study on Influenza in US. NDSSL; 2015.
9. Hall IM, Gani R, Hughes HE, Leach S. Real-Time Epidemic Forecasting for Pandemic Influenza. Epidemiology and Infection. 2007;135:372–385.
10. Lampos V, Cristianini N. Nowcasting Events from the Social Web with Statistical Learning. ACM Transactions on Intelligent Systems and Technology (TIST). 2012;3(4):72.
11. Preis T, Moat HS. Adaptive Nowcasting of Influenza Outbreaks using Google Searches. Royal Society Open Science. 2014;1(2).
12. Brooks L, Hyun S, Tibshirani R, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Computational Biology. 2015;11(8).
13. McIver DJ, Brownstein JS. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol. 2014;10(4):e1003581.
14. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013–2014 Influenza Season Using Wikipedia. PLoS Comput Biol. 2015;11(5):e1004239.
15. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014;343(14 March).