Proceedings of the ASME 2018
International Mechanical Engineering Congress and Exposition
IMECE2018
November 9-15, 2018, Pittsburgh, PA, USA

IMECE2018-86559

THE WISDOM OF CROWDS APPROACH TO INFLUENZA-RATE FORECASTING

Jeffrey J. Morgan[1]
Catholic University of America
Washington, DC, USA

Otto C. Wilson
Catholic University of America
Washington, DC, USA

Prahlad G. Menon
University of Pittsburgh
Pittsburgh, PA, USA

ABSTRACT

Influenza is an important public health concern. Influenza leads to the death or hospitalization of thousands of people around the globe every year. However, the flu-season varies every year viz. when it starts, when it peaks, and the severity of the outbreak. Knowing the trajectory of the epidemic outbreak is important for taking appropriate mitigation strategies. Starting with the 2013-2014 flu season, the Influenza Division of the Centers for Disease Control and Prevention (CDC) has held a "Predict the Influenza Season Challenge" to encourage the scientific community to make advances in the field of influenza forecasting. A key observation from these challenges is that a simple average of the submitted forecasts outperformed nearly all of the individual models. Further, ongoing efforts seek ways to assign weights to individual models to create high-performing ensemble models. Given the sheer number of models, as well as variation in methodology followed among teams contributing influenza-risk forecasts, multiple forecasting models can be combined, by capturing human judgment, to outperform a simple average of these same models. This project exploits such a "wisdom of crowds" approach, using public votes acquired with the help of an R/Shiny based web-application platform in order to assign weights to individual forecasting models on a week-over-week basis, in an effort to improve overall ILI risk prediction accuracy. We describe a strategy for improving the accuracy of influenza risk forecast modeling based on a crowd-sourced set of team-specific forecast votes and the results of the 2017-2018 season. Our approach to assigning weights based on crowd-sourced votes on individual models outperformed an average forecasts of the individual models. The crowd was statistically significantly more accurate than the average model and all but one of the individual models.

1.0 INTRODUCTION

Influenza has long been an important public health concern. 2018 marks the 100th anniversary of the 1918 Spanish flu pandemic. Estimates vary, but it is thought to have infected 500 million people and killed 20 to 50 million people worldwide [1]. While influenza pandemics occur infrequently, seasonal influenza presents a global health burden that varies in timing and intensity every year; millions become ill and thousands die every year [2]. Many factors, including characteristics of the circulating virus, vaccine effectiveness, weather, and human behavior, contribute to the variations in the season. The Public Health community has multiple non-pharmaceutical intervention (NPI) strategies to reduce influenza, including public service announcements (e.g., encourage vaccinations, hand washing, covering coughs and sneezes), social distancing (e.g., school closures), and environmental surface cleaning and disinfecting (e.g., frequently touched surfaces in schools and airplanes) [3]. To encourage the scientific community to collaborate and make advances in the field of infectious disease forecasting, CDC's Influenza Division has held "Predict the Influenza Season Challenges" since the 2013-2014 flu season. A key observation from these previous challenges is that a simple average of the submitted forecasts outperformed nearly all of the individual models. A simple average with equal weights to all models demonstrated relatively good performance, but a weighting strategy that assigns more weight to better models offers an opportunity for improved performance. This paper explores a proof of principle approach of combining models by assigning weights based on votes collected from a crowd of willing participants.

[1] Jeffrey Morgan is also a scientist with Joint Research and Development, Inc.

2.0 METHODS

The CDC compiles data about the fraction of influenza-related patient visits across the US and its territories, from ~2,000 healthcare providers. CDC publishes weekly data on current and historical flu seasons on their FluView website [4]. The website is updated every Friday, around noon, and is freely available to the general public.

2.1 CDC Influenza Metrics

The CDC includes weighted Influenza Like Illness (wILI) as a metric for influenza activity. The CDC derives wILI as the percentage of people attending participating health facilities with symptoms that suggest an influenza like illness, weighted based on population [5]. Table 1 shows the results from a CDC Study of the influenza seasons in the U.S. from 2010-2011 to 2015-2016 and includes their relative ranking. Table 1 shows that the clear relationship between wILI and other measures of influenza burden.

Table 1. Influenza Burden in the U.S. from the 2010-2011 Season through the 2015-2016 Season and their Relative Ranking [6].

	Esitmated Illnesses	*Estimated Hospit-alizations*	*Estimated Deaths*	*Peak wILI%*	*Mean wILI% (weeks 40-39)*
2010-2011	21,096,749 (5^{TH})	281,589 (5^{TH})	13,541 (4^{TH})	4.552 (4^{TH})	1.766 (3^{RD})
2011-2012	9,231,004 (6^{TH})	139,497 (6^{TH})	4,154 (6^{TH})	2.389 (6^{TH})	1.447 (6^{TH})
2012-2013	35,490,424 (1^{ST})	592,688 (2^{ND})	19,962 (1^{ST})	6.061 (1^{ST})	1.905 (2^{ND})
2013-2014	28,445,377 (3^{RD})	322,123 (3^{RD})	13,590 (3^{RD})	4.591 (3^{RD})	1.738 (4^{TH})
2014-2015	34,292,299 (2^{ND})	707,155 (1^{ST})	19,490 (2^{ND})	5.982 (2^{ND})	1.975 (1^{ST})
2015-2016	24,577,163 (4^{TH})	308,232 (4^{TH})	11,995 (5^{TH})	3.560 (5^{TH})	1.652 (5^{TH})

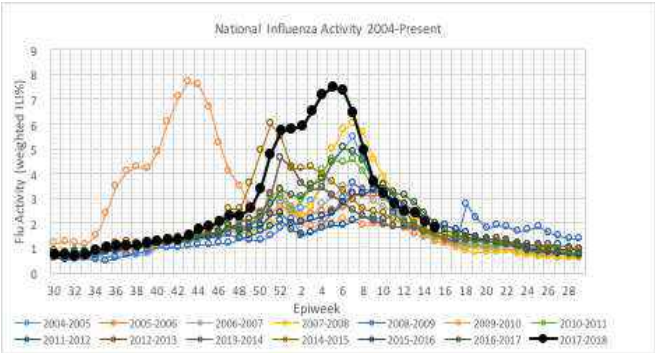

Figure 1. *Epidemiological Curve for Influenza at the National level from 2004-present [4].*

The rise and fall of wILI over the course of these seasons is shown in the epidemiological curve of Figure 1. Clearly, some seasons have higher peaks than others, but further inspection shows that there is variation in the timing for the season peak and season onset. Season offset for a region is defined as the first of three consecutive weeks above baseline for that region. The baseline is based on the wILI during the weeks outside of the influenza season; it is the mean wILI plus two standard deviations [5]. As shown in Figure 2, the U.S. Department of Health & Human Services (HHS) groups states and territories within the U.S. into ten different regions. The baseline wILI% for each region are shown in Table 2.

Figure 2. Map of HHS Regions

Table 2. HHS Regions and their Influenza Baseline for the 2017-2018 Season [5]

Region	*States*	*2017-21018 Baseline*
Region 1	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont	1.4%
Region 2	New Jersey, New York, Puerto Rico, and the U.S. Virgin Islands	3.1%
Region 3	Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia	2.0%
Region 4	Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee	1.9%
Region 5	Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin	1.8%
Region 6	Arkansas, Louisiana, New Mexico, Oklahoma, and Texas	4.2%
Region 7	Iowa, Kansas, Missouri, and Nebraska	1.9%
Region 8	Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming	1.3%
Region 9	Arizona, California, Hawaii, and Nevada	2.4%
Region 10	Alaska, Idaho, Oregon and Washington	1.4%

2.2 CDC Influenza Challenge

With the ultimate goal of improving the prevention and control of influenza, CDC launched the influenza forecasting challenge in 2013-2014. The stated short-term goals were to improve the understanding of influenza models and their contribution to decisions in public health [7]. CDC has continued to conduct influenza challenges and to incorporate lessons learned by the modelers, influenza challenge organizers, and

CDC Influenza Division personnel who are briefed on model results.

On a weekly basis throughout the challenge, modeling teams incorporate the CDC data that is released on the FluSight website on Fridays and submit forecasts on Monday. CDC uploads these models to GitHub on Tuesdays. For the 2017-2018 season, forecasters provided discrete probability distributions for the seven targets in Table 3 at the National level and each of the ten HHS Regions described previously [8]. Each increment in the discrete probability distribution can be thought of as a "bin".

Table 3. Influenza Challenge Targets and Basis for Scoring

Target	*Definition*	*Scoring Range for Forecasting Challenge*
Season Onset Week	The first of 3 consecutive weeks that wILI% is above baseline	Actual week ± 1 week
Season Peak Week	The week that the wILI% is highest	Actual week ± 1 week
Season Peak Intensity (truncated to nearest 0.1%)	The wILI% of the Peak Week	wILI% ± 0.5%
1 week ahead wILI% (truncated to nearest 0.1%)	The wILI% of the week 1 week in advance (the next week that CDC will provide data for and also the week that just concluded)	wILI% ± 0.5%
2 week ahead wILI% (truncated to nearest 0.1%)	The wILI% of the week 2 weeks in advance	wILI% ± 0.5%
3 weeks ahead wILI% (truncated to nearest 0.1%)	The wILI% of the week 3 weeks in advance	wILI% ± 0.5%
4 weeks ahead wILI% (truncated to nearest 0.1%)	The wILI% of the week 4 weeks in advance	wILI% ± 0.5%

For targets based on weeks (i.e., season onset and season peak week), the probability distribution bins are weeks. For these targets, scoring is based on the probability assigned to the bin containing the correct week as well as one bin (aka one week) before and one bin after. For targets based on wILI%, the probability distribution "bins" are in 0.1% increments and are truncated such that 3.01 and 3.99 both fall within the bio of 3.0 to 3.1. For these wILI% targets, points are awarded for the probability assigned to the bin containing the correct wILI% as well as five bins (aka 0.5%) below and five bins (aka 0.5%) above. All bins within the scoring range yield equal scores (i.e., bonus points are not given for a prediction within the exact bin).

When analyzing the results of previous challenges, CDC found that a simple mean of the forecasts outperforms nearly all of the individual models. There are many intutitive ways that one might improve upon an ensemble model which simply takes the mean of all the models. For example, one could give more weight to better performing models and less weight to models with less impressive track records. Computer algorithms could be written to assess various weighting strategies based on historic performance or on recent performance. For the 2017-2018 season, several efforts were undertaken to explore various strategies to combine models [9].

For the CDC Influenza Challenge, teams incorporate Friday's FluView data into their models and submit forecasts to CDC by Monday night. CDC uploads these models on GitHub [8], on Tuesday morning or early afternoon.

It is important to note that there is a lag between when a patient presents at a health care center and when that visit is incorporated into totals released by CDC on Friday afternoon. At the end of every week, participating health providers submit their data to their state health departments. The state health departments compile that data and submit to CDC during the early part of the following week. CDC compiles that data and releases that data on Friday afternoon. As a result, if a patient visits a health center on Monday of week *n*, that visit will be reflected in the data shared by CDC on Friday of week *n+1*.

2.3 Wisdom of Crowds Approach

Human brains are very powerful and process tremendous amounts of data at the subconscious level. For example, we can be outdoors and gain a sense that a storm is coming even when we had no conscious thought to consider the likelihood of a storm – going through a predetermined checklist of determining the type of clouds in the sky or estimating the speed and direction of the wind, or assessing the change in barometric pressure. We have developed a sense of an impending storm without using a well-defined algorithm with threshold values to make this assessment. Somehow, we gain this experience and place some amount of trust in our predictions without making any quantitative calculations. We accept that our predictions are not always right, but we continue to make predictions. This "Wisdom of Crowds" approach aims to capitalize in our innate abilities to assess the trajectory of the influenza season.

Outputs from various influenza models may show different disease trajectories and then be given to a leader to aid in the decision making process. To synthesize these forecasts, the decision maker may want to know how well each model has performed historically and recently. Decision makers may incorporate their own judgment of the relative merits of each metric and may combine that information with their own mental model to formulate the relative likelihood of various outcomes. However, the decision maker may not be particularly adept at mentally combining forecasts. The book, Superforecasting The Art and Science of Prediction by Phillip Tetlock and Dan Gardener notes that teams are better at forecasting than individuals and that crowds are more accurate than professional forecasters. Through the Good Judgement Project, the Intelligence Activity Research Projects Activity (IARPA) discovered that there are a few people who are really good at making predictions and coined those people "Superforecasters". Tetlock and Gardener describe characteristics of Superforecasters and noted that the traits of superforecasters conflict with traits that make good leaders [10].

The book, The Wisdom of Crowds, by James Surowiecki, explains the phenomena that crowds are often able to derive correct solutions even if the judgement of individual

crowdmemebers is unexceptional. For example, if a group of people guess the number of jelly beans in a glass jar, many guesses will be incorrect by wide margins, but the mean may be fairly accurate. One interesting aspect of this phenomena is that the crowd participants don't have to be experts. In fact, it is better if there is a broad background amongst the crowd members [11].

2.3 Wisdom of Crowd Members

Organizers of CDC's Influenza Challenge have provided briefs of this challenge to the Office of Science and Technology Policy Pandemic Prediction and Forecasting Science and Technology Working Group (PPFST WG). These organizers are regular participants in the PPFST WG and include one of the PPFST WG co-chairs. The idea for this project – to assemble a crowd of interested people who would vote on the models they thought would score the most points - was discussed with Influenza Challenge leaders, Roni Rosenfeld, who leads the Delphi Group at Carnegie Mellon University, Nicholas Reich who is leading several efforts in creating ensemble models, and the PPFST WG. Crowd members whom were invited to vote on their top-model to believe in on a given week, were identified from the PPFST WG, other modeling teams and select individuals whom were familiar with the ILI forecasting project.

2.4 Models

It was deemed important to identify good models for the crowd members to select from. A model from Columbia University (CU) won the inaugural Influenza Challenge [7]. Models from Carnegie Mellon's Delphi Group have won the subsequent three challenges [12], with models from Columbia University also placing high. The top model that was not developed by the Delphi Group in 2015-2016 was Kernel of Truth from University of Massachusetts – Amherst. Similarly, the top model not developed by the Delphi Group or Columbia University in 2016-2017 was developed by Los Alamos National Laboratory (LANL). Two other models were provided to the crowd. Models from Knowledge Based Systems Inc (KBSI), which is affiliated with Texas A&M University, and 4Sight, from the Biocomplexity Institute of Virginia Tech, participated in both the 2015-2016 challenge and the 2016-2017 challenge, with demonstrated improvement. KBSI has been awarded Phase I and Phase II Small Business Innovative Research (SBIR) awards for the Data Integration and Predictive Analysis System (IPAS) which they used in these challenges [13]. Virginia Tech developed an agent-based model for the Ebola 2014 outbreak which subsequently had a strong performance in the Ebola Challenge organized by the Research and Policy for Infectious Disease Dynamics (RAPIDD) at the National Institutes of Health (NIH) [14]. The Delphi Epicast model and the 4Sight models incorporate crowd sourcing in their models. This paragraph described the basis for which the models for the "Wisdom of Crowds" effort were selected.

With the aim of presenting a manageable number of choices, crowd members were only given a few models from which to choose for the first few weeks of the challenge. At the request of participants, more models were made available. Following Epidemic Week (EW) 5 of 2018, the 4Sight model was no longer submitted. As a result, 4Sight2 was substituted as an option in its place. Table 4 shows the availability of models, as well as the number of crowd members who submitted votes for each week.

Table 4. Number of Participating Crowd Members and Models to choose from for the MWWR Epidemiological Weeks.

Epi Week	# in crowd	Delphi Epicast	Delphi STAT	CU-SEL	CU-Network	KBSI	KOT	LANL DBM	4Sight	4Sight2
43	3		X					X		
44	2		X	X			X			
45	1*	X	X	X	X	X				
46	4	X	X	X	X	X				
47	5	X	X	X	X	X	X			
48	5	X	X	X	X	X	X			
49	3	X	X	X	X	X	X			
50	4	X	X	X	X	X	X	X		
51	3	X	X	X	X	X	X	X		
52	3	X	X	X	X	X	X	X	X	
1	3	X	X	X	X	X	X	X	X	
2	3	X	X	X	X	X	X	X	X	
3	3	X	X	X	X	X	X	X	X	
4	4	X	X	X	X	X	X	X	X	
5	4	X	X	X	X	X	X	X	X	
6	3	X	X	X	X	X	X	X		X
7	4	X	X	X	X	X	X	X		X
8	4	X	X	X	X	X	X	X		X
9	3	X	X	X	X	X	X	X		X
10	3	X	X	X	X	X	X	X		X
11	4	X	X	X	X	X	X	X		X
12	4	X	X	X	X	X	X	X		X
13	4	X	X	X	X	X	X	X		X
14	4	X	X	X	X	X	X	X		X
15	4	X	X	X	X	X	X	X		X

*There was only one participant, who submitted two ballots, for EW 45 which required picks to be made on Thanksgiving Day

2.4 R/Shiny App

An R/Shiny app was created to download the models from GitHub, display the models, and provide a means to capture and record votes from crowd member [15]. When the user submits a vote for a particular target, it automatically advances to the next target. This app also includes a tab with epidemiological curves from previous seasons and another tab showing the performance of the models in the challenge. All of this information is considered potentially useful in deciding which model is likely to be most accurate for a particular target and location. The app also calculates the score, as described previously, for each of the predictions from each model. The R/Shiny app is hosted on a DigitalOcean. server. A screenshot of the R/Shiny app is shown in Figure 3.

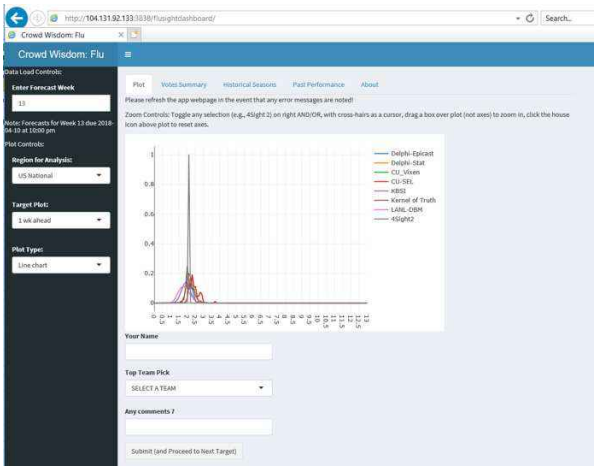

Figure 3. Screenshot of R/Shiny app for Influenza Forecasting

2.4 Wisdom of Crowds Mechanics

Except for holiday weeks, CDC posted models some time on Tuesday morning or early afternoon. (For holiday weeks, the timing was adjusted to allow for changes in CDC reporting requirements.) The models and most recent CDC wILI data were subsequently uploaded to the Rshiny-based website for capturing crowd votes [15] as technical difficulties permitted. Crowd members were also emailed a ballot and optional worksheet that included epidemic trajectories with the wILI% for the current season and all seasons since the 2004-2005 season. The worksheet provided a means for participants to record their thoughts on each of the targets while looking at the epidemic curves prior to selecting from the available models.

Crowd members were asked to vote on the website (or via the ballot/email) for the seven targets of Table 3, at the National Level, HHS Region 3 and HHS Region 4. In the interest of time crowd members were provided an option to allow their votes for best model for each of the targets at the national level to count for HHS Regions 3 and 4 and also the opportunity to allow their votes for HHS Region 3 to count for HHS Region 4. The CDC challenge required votes for all ten HHS Regions. Votes for HHS Regions 3 and 4 were averaged and used to determine weights for HHS Regions 1, 2, and 5-10. (Voting for HHS Regions 3 and 4 was adapted beginning in the 3rd week of the challenge (EW 45). For the first week (EW 43), crowd members only voted on HHS Region 1. For the second week (EW 44), crowd members voted on HHS Regions 1 and 2.) The results from Regions 1, 2, and 5-10 have not yet been assessed and are not reported here.

Throughout the season, crowd members were provided feedback on the accuracy scores for each of the models, as well as their individual selections. When crowd members were provided updates of the scores, the top two leading crowd members for that week were identified.

3.0 RESULTS

3.1 Overall Results

Data for the proposed wisdom-of-crowds' approach is emerging as the influenza challenge has not yet reached its completion. (This section will be updated after the challenge has reached its completion and data has been analyzed). Results for the PPFST Crowd prediction is compared to the models that were offered to crowd members for that week, as not all models were available for every week. Results of the crowd viz. average and crowd viz. each model is shown for each week is shown in and each target in Tables XX and XY, respectively.

Current data collected from this approach thus far indicates that the mean accuracy score of the crowd-selected team on a weekly basis across territories and regions in the study was consistently higher than the average ILI prediction accuracy score across contributing teams on a weekly basis ($p<0.01$) and per target ($p<0.01$, see Figure 4). The median difference of means between score assigned to the crowd as opposed to the average score across teams was 0.02 higher in favor of the crowd, whereas the mean difference between individual teams was as high as 0.26 when compared against individual team-specific average scores, across regions and targets for prediction (see Table 6).

Table 5. Weekly performance of Crowd and each of the models

weekc	Delphi Epicast	Delphi STAT	CU-SEL	CU-Network	KBSI	Kernel of Truth	LANL DBM	4Sight	4Sight 2	Crowd	AVG CHOICE
EW 43 11-06-17	0.558	0.509	0.407	0.411	0.515	0.502	0.432	0.714	0.561	0.471	0.502
EW 44 11-13-17	0.502	0.465	0.444	0.420	0.394	0.413	0.435	0.703	0.682	0.472	0.426
EW 45 11-20-17	0.733	0.691	0.475	0.362	0.597	0.376	0.558	0.920	0.910	0.527	0.571
EW 46 11-30-17	0.472	0.509	0.355	0.312	0.476	0.430	0.454	0.693	0.677	0.424	0.425
EW 47 12-05-17	0.428	0.405	0.388	0.383	0.448	0.368	0.115	0.519	0.432	0.434	0.403
EW 48 12-12-17	0.392	0.433	0.384	0.393	0.369	0.343	0.373	0.400	0.408	0.425	0.386
EW 49 12-19-17	0.388	0.400	0.383	0.384	0.350	0.304	0.358	0.275	0.312	0.397	0.368
EW 50 12-28-17	0.361	0.314	0.313	0.295	0.300	0.270	0.291	0.212	0.225	0.346	0.306
EW 51 01-05-18	0.324	0.290	0.299	0.302	0.330	0.266	0.253	0.184	0.472	0.318	0.281
EW 52 01-09-18	0.277	0.361	0.281	0.266	0.342	0.288	0.411	0.202	0.281	0.304	0.303
EW 01 01-17-18	0.266	0.332	0.247	0.240	0.333	0.247	0.392	0.183	0.454	0.299	0.280
EW 02 01-24-18	0.241	0.294	0.214	0.205	0.291	0.216	0.345	0.155	0.454	0.251	0.245
EW 03 01-30-18	0.297	0.293	0.311	0.282	0.244	0.224	0.370	0.147	0.455	0.314	0.271
EW 04 02-06-18	0.521	0.341	0.500	0.503	0.365	0.372	0.391	0.232	0.360	0.460	0.403
EW 05 02-13-18	0.483	0.387	0.428	0.522	0.513	0.513	0.397	0.355	0.360	0.472	0.450
EW 06 02-20-18	0.424	0.430	0.407	0.477	0.568	0.495	0.368	0.367	0.312	0.416	0.435
EW 07 02-27-18	0.561	0.609	0.565	0.559	0.569	0.574	0.499	0.000	0.444	0.576	0.550
EW 08 03-06-18	0.662	0.578	0.632	0.640	0.533	0.595	0.542	0.000	0.667	0.618	0.606
EW 09 03-13-18	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.750	0.659
EW10 03-20-18	0.894	0.760	0.846	0.824	0.539	0.750	0.752	0.000	0.476	0.823	0.730
EW11 03-27-18	0.882	0.806	0.868	0.845	0.536	0.795	0.798	0.000	0.714	0.873	0.780
EW12 04-03-18	0.896	0.807	0.830	0.826	0.574	0.832	0.816	0.000	0.762	0.852	0.793
EW13 04-10-18	0.902	0.836	0.869	0.870	0.778	0.866	0.887	0.000	0.933	0.884	0.868
EW14 04-17-18	0.863	0.772	0.775	0.782	0.768	0.825	0.895	0.000	0.867	0.821	0.818
EW15 04-24-18	0.923	0.902	0.928	0.940	0.765	0.931	0.949	0.000	1.000	0.938	0.917

Each team produces a forecast with a probability distribution of forecasted values for each target in each region. As discussed in Section 2.2, the accuracy score for the purpose of this analysis and the CDC challenge is defined as the probability assigned to the actual correct ILI risk estimate for the given week (once the week has elapsed) +/- the average across a range of values on the aforementioned probability distribution, which is +/- 1 week for temporal-evolution of peak-ILI risk as well as season-onset predictions, or +/- 0.5% for ILI predictions for 1-to-4 weeks ahead.

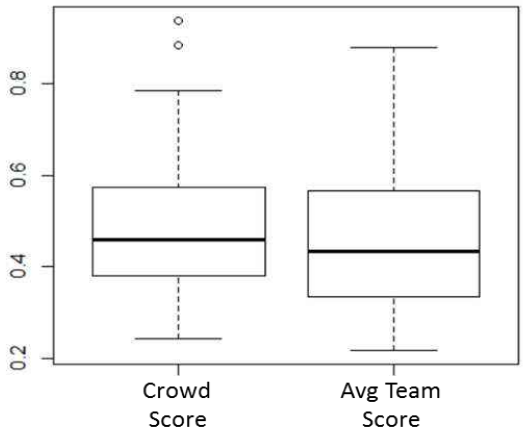

Figure 4. Box Plot of Crowd Score and Average Team Score

4.0 DISCUSSION

It is important to note that the results are tentative. Tables 5 and 6 assume that the influenza season has peaked at the week and wILI% most recently reported by the CDC. Every week, CDC updates their FluSight website with influenza totals to date. Some local health providers do not report their results on a timely basis. A health provider that is overwhelmed with influenza cases may not have the ability to accurately break down and report the number of influenza cases. When the data is reported at a later date, the additional information affects the CDC totals. Every week, the CDC reported values change such that prior scores are affected. Even slight changes to wILI% (e.g., 4.03 to 3.98) can cause a shift in the scoring. This effect is exacerbated for the 4Sight2 model which often responded with a point prediction with a probability of 1.000 in a single bin. While this may turn out to be a good strategy to maximize points, a change to wILI% such that their point prediction falls out of the scoring range will significantly alter the points scored for that region.

Table 6. Comparison of Individual Models, Crowd, and Average

MEAN DIFFERENCE ACROSS TARGETS / LOCATIONS	P-VALUE	CI LOWER	CI UPPER	TEAM V/S CROWD
-0.02	0.00	-0.04	-0.01	Delphi Epicast
0.02	0.01	0.01	0.03	Delphi STAT
0.02	0.02	0.00	0.04	CU-SEL
0.02	0.05	0.00	0.05	CU-Network
0.06	0.00	0.02	0.09	KBSI
0.04	0.00	0.02	0.06	Kernel of Truth
0.03	0.05	0.00	0.07	LANL DBM
0.26	0.00	0.19	0.32	4Sight
0.00	0.82	-0.03	0.04	4Sight2
0.03	0.00	0.02	0.04	Average of Teams

As Table 5 shows, performance increased in the latter half of the challenge. There are several explanations for this observation. The results for several targets have already been tentatively identified and the general shape of the epidemiological curve is flattening. After CDC reports influenza activity with three consecutive weeks with wILI% above baseline, the season onset is known. Once the season peak has been reached and the epidemiological curve has turned downwards, the season peak week and peak percentage are tentatively known. These seemingly known results are subject to fluctuations in the data reported by the CDC, as described previously.

Of course, there is some probability that there will be an influenza wave late in the season which will create an even later peak with higher values. Only one of the eight models, KBSI, assigns much probability (0.25 as of EW15) to the likelihood of a late influenza wave exceeding earlier peaks. One crowd member reflected this belief in their votes. Since the tentative scoring in this analysis considers the peak to have passed, KBSI appears to be less accurate in predicting the peak. Since one crowd member regularly selects KBSI for this target, the crowd score is also lower than it would be otherwise. While there is some probability that there will be a late influenza wave, models attributing 0.0 probability to this happening will normally score lower. This does not mean that attributing 0.0 probability to a late influenza wave is more correct than one that does not. Hopefully, with a large enough crowd, these low probability events will be appropriately accounted for with an appropriate percentage of votes.

As shown in Table 4, most weeks only had three or four participants. As a result, if one participant chose a model that performed poorly, the crowd's performance would be lessened.

The study was generally successful in showing at a "proof-of-concept-level" that a crowd can be solicited to choose models on a weekly basis; the crowd's input can be consolidated into a "Crowd Vote"; and that the "Crowd Vote" may outperform a simple average of the available models. However, a larger crowd size is needed to fully assess the benefits of a wisdom of crowds approach.

After the first several weeks, crowd members were able to choose from 8 models. It is not clear what is the optimal number of choices. Studies accessed prior to the start of this "Wisdom of Crowds" challenge suggested a "less is more approach". Research by Sheena Iyengar of the Columbia Business School has indicated that three options is often the ideal number of choices. For other choices, her findings suggest that three columns with three options each would be ideal. Simply, too many choices can be overwhelming [16]. These authors have not yet found a study on the optimal number of choices for discrete probability functions.

There are several possible reasons why people who were aware of the challenge did not participate. Many people are busy with other time consuming and worthwhile efforts. Selecting the best model was a somewhat time-consuming process, several times longer than the 5 minutes which was originally targeted. The task asked for crowd members to make choices for 7 targets for three locations (National, HHS Region 3, and HHS Region 4). Perhaps, 21 selections with 8 choices is overwhelming.

While the process allowed for crowd members to use model selection choices from one location to be used for another location, this option was only exercised twice in the 87 ballots submitted to date. It is interesting to note that the Delphi Epicast model asks for participants to "sketch" their expectations for the epidemiological curve for 21 16 locations (National, ten HHS Regions, four states, and Washington, D.C.). Other people may not have participated because they are uncomfortable or unfamiliar with probability distribution functions and unsure how to compare their expectations for the trajectory of the epidemiological curve with the discrete probability distributions of the 8 models. Others may have questioned whether their experience was appropriate for this study; they may feel that they have either "too much experience" or "too little experience" which would skew the results one way or the other. Others may have declined to participate because they perceived the process to be not enjoyable or not worthwhile.

The crowd members reflected a "coalition of the willing". No prizes for participating or winning were announced. The intent of offering a prize would be to achieve a larger crowd and, more importantly, better results. Prize money would provide motivation for a crowd member to participate every week and would provide motivation for making wise selections. On the hand, the desire to win, which could be enhanced with the opportunity for prize money, may drive individuals to gambel aggressively and select the 4Sight2 model which offers the opportunity to score a 1.0. The Delphi Epicast model output, on the other hand, always follows a discrete Gaussian distribution and offers a more conservative approach and predictable return. If prize money were provided to the crowd member with the overall highest score for the season, that could drive a crowd member who was trailing to select models which provide an opportunity to catch up. For example, selecting KBSI, which assigned more likelihood to a late influenza wave, for the season peak target to give them a chance to catch up.

5.0 CONCLUSION

In this study, using a crowd to identify the models most likely to be correct resulted in more accurate forecasts, as defined by the study, than a simple average. Due to the nature of the variability in the influenza season and the members of the crowd, future studies may yield different results. Further study is warranted

The authors have begun planning for the 2018-2019 influenza challenge. The authors intend to recruit more crowd members for the next season and conduct the challenge with more rigor. The PPFST WG will remain a focus area for participation, but crowd members may be recruited from universities and social media.

A prize structure for Wisdom of Crowd participants will also be determined and announced prior to next season's challenge.
In future, we propose to correlate the nature i.e. time-series pattern of ILI risk increase after season onset, in a given season, in addition to descriptive statistics on voted-team-specific ILI risk accuracy scores.

The authors will explore stratification of crowd voters by experience level with technology and Flusight dataset (i.e. low, medium, high) and then regress their voting scores on their experience level.

ACKNOWLEDGMENTS

The authors would like to thank the following individuals for their contributions as voting members of the crowd in this study: Emily Briskin, Jessica Deerin, Bryan Lewis, Paul Lewis, Steven Morgan, Jonathan Phillips, Srinivasan Venkatramanan, and Cécile Viboud.

The authors would like to thank the following individuals for guidance in developing and executing this study: Matt Biggerstaff, Jean-Paul Chretien, Frederick Dahlgren, Robert Huffman, Michael Johannson, Paul Lewis, Craig McGowan, Nicholas Reich and Roni Rosenfeld.

Jeffrey Morgan serves as Executive Secretary of Prediction and Forecasting Science and Technology (PPFST) and lead various Focus Areas. He also provided contractor support to the government (COR) who managed the KBSI SBIR.

REFERENCES

[1] History Channel. "Spanish Flu". https://www.history.com/topics/1918-flu-pandemic
[2] Seasonal Influenza, More Information. https://www.cdc.gov/flu/about/qa/disease.htm
[3] Qualls, Noreen et.al. Morbidity and Mortality Weekly Report (MWWR) "Community Mitigation Guidelines to Prevent Pandemic Influenza – United States, 2017". April 21, 2017, 66(1);1-34.
[4] CDC FluView: https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html
[5] CDC Overview of Influenza Surveillance in the United States: https://www.cdc.gov/flu/weekly/overview.htm
[6] CDC Estimated Influenza Illnesses, Medical Visits, Hopitalizations, and Deaths Avervted by Vaccination in the United States.
[7] Biggerstaff, et.al. BMC Infectious Diseases. "Results from the centers for disease control and prevention's predict the 2013-2014 Influenza Season Challenge". 16:357. 22 July 2016.
[8] GitHub cdcepi/FluSight-forecasts: https://github.com/cdcepi/FluSight-forecasts
[9] Biocompare. "Forecasting Influenza in U.S. Is a Group Effort" posted December 28, 2017.
[10] Tetlock, Philip and Gardner, Dan. Superforecasting The Art and Science of Prediction. Crown Publishers. New York: 2015.
[11] Surowiecki, James. The Wisdom of Crowds. Doubleday. New York: 2004.
[12] DELPHI: Developing the Theory and Practice of Epidemiological Forecasting https://delphi.midas.cs.cmu.edu/
[13] SBIR STTR Data Integration and Predictive Analysis Systems (IPAS).

[14] Venkatramanan, Srinivasan, et.al. Epidemics. “Using data-driven agent-based models for forecasting emerging infectious diseases”. Epidemics22(2018)43-49. 22 February 2017.
[15] http://104.131.92.133:3838/flusightdashboard/
[16] Chambers, Kipp. “Number of Choices in Survey Questions: How Much is Too Much?” surveygizmo//resources. Dec 1, 2011