Real-time Estimation of Disease Activity in Emerging Outbreaks using Internet Search Information

Emily L. Aiken[1, *], Sarah F. McGough[2], Maimuna S. Majumder[3], Gal Wachtel[4], Andre T. Nguyen[5, 6], Cecile Viboud[7], Mauricio Santillana[1, 4, 8, *]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138
[2]Harvard T.H. Chan School of Public Health, Boston, MA 02115
[3]Department of Healthcare Policy, Harvard Medical School, Boston, MA 02115
[4]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02215
[5]Booz Allen Hamilton, Columbia, MD 21044
[6]University of Maryland, Baltimore County, Baltimore, MD 21250
[7]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892
[8]Department of Pediatrics, Harvard Medical School, Boston, MA 02215
[*]**Correspondence to:** Emily Aiken <emilyaiken@berkeley.edu> and Mauricio Santillana <msantill@g.harvard.edu>

Abstract

Understanding the behavior of emerging disease outbreaks in, or ahead of, real-time could help healthcare officials better design interventions to mitigate impacts on affected populations. Most healthcare-based disease surveillance systems, however, have significant inherent reporting delays due to data collection, aggregation, and distribution processes. Recent work has shown that machine learning methods leveraging a combination of traditionally collected epidemiological information and novel Internet-based data sources, such as disease-related Internet search activity, can produce meaningful "nowcasts" of disease incidence ahead of healthcare-based estimates, with most successful case studies focusing on endemic and seasonal diseases such as influenza and dengue. Here, we apply similar computational methods to emerging outbreaks in geographic regions where no historical presence of the disease of interest has been observed. By combining limited available historical epidemiological data available with disease-related Internet search activity, we retrospectively estimate disease activity in five recent outbreaks weeks ahead of traditional surveillance methods. We find that the proposed computational methods frequently provide useful real-time incidence estimates that can help fill temporal data gaps resulting from surveillance reporting delays. However, the proposed methods are limited by issues of sample bias and skew in search query volumes, perhaps as a result of media coverage.

Introduction

Disease outbreaks have been major drivers of morbidity and mortality since the beginning of recorded history and continue to pose a major threat to humankind. Surveillance of disease out-

breaks by healthcare systems is key to effective outbreak response. In particular, surveillance data is necessary to determine the overall scale of response to an outbreak, allocate limited resources for treatment and prevention, and effectively time interventions to minimize impacts [1]. Epidemiologists use surveillance data to estimate important features of an outbreak, such as morbidity and mortality burden, case fatality rate, and transmission patterns. In recent years, the use of mathematical modeling of disease activity and transmission to predict the likely trajectory of an outbreak and guide intervention strategies has been increasingly explored [1, 2, 3, 4].

It is particularly challenging to monitor and characterize unexpected (emerging) disease outbreaks in regions that have not experienced the presence of a specific pathogen in recent times. Such emerging disease outbreaks, particularly in their early stages, are characterized by incomplete, delayed, and biased epidemiological surveillance data [1]. Reporting delays in surveillance systems inevitably emerge from limited healthcare resources and coverage, as well as the time required to process lab tests and clean, anonymize, aggregate, and communicate data from distributed healthcare facilities to central authorities. These reporting delays and issues of missingness are manifested in epidemiological reports released by the World Health Organization (WHO) and other health authorities for several recent outbreaks [5, 6, 7, 8, 9, 10, 11].

Novel Internet-based data sources have the potential to fill some of these temporal "data gaps" in tracking emerging outbreaks. Research to date on using Internet-based data sources to provide early estimations of disease activity has shown promising results for endemic diseases in high- and middle-income countries, including influenza in the United States [12, 13, 14, 15, 16, 17] and dengue in Brazil, Mexico, Thailand, Singapore, and Taiwan [18]. Digital epidemiological methods use mathematical methods to combine Internet-based data – including Google search trends (data on aggregated Google query volumes) [12, 13, 18], Twitter microblogs [14, 15, 19], online news aggregators [20], electronic medical records [21, 22], and crowdsourced disease activity estimates [23, 24] – with historic epidemiological data to produce real-time estimates of disease activity ("nowcasts").

One particularly well-studied method for tracking seasonal and endemic diseases is ARGO, a machine learning approach based on a dynamic multivariate regularized regression that leverages historic epidemiological data along with real-time generalized online data sources, including Google search trends, Twitter microblogs, electronic health records, and others [13, 25]. ARGO has been shown to produce meaningful and accurate national-level disease activity estimates for influenza in the US and Latin America, and dengue in several middle income countries weeks ahead of reports issued by traditional surveillance systems [13, 17, 18].

Adapting digital epidemiological methods like ARGO for tracking emerging outbreaks brings up a host of new challenges relating to an absence of historical epidemiological data for training and validation, and a paucity of digital data due to poorer internet coverage. To our knowledge, three past studies have experimented with Internet-based data for emerging infections: Majumder et al. [26] demonstrate the use of digital data sources (including Google search trends and news reports) to provide estimates of R_0, the basic reproductive number, in the absence of real-time epidemiological surveillance data in the 2016 Latin American Zika outbreak. Chunara et al. [27] use Twitter and news report data to estimate R_0 in the 2010 Haitian Cholera outbreak. In the only work to date on nowcasting disease incidence in an emerging outbreak with digital data sources, McGough et al. [28] incorporate information from Google search trends, Twitter, and news reports to produce

accurate nowcasts of incidence in the 2015-2016 Latin American Zika outbreak 1-3 weeks in advance of standard epidemiological reports.

Our Contribution. Here we expand on [28] to evaluate the performance of digital epidemiological methods for nowcasting five contemporary outbreaks: Yellow Fever in Angola (2016), Zika in Colombia (2015-2016), Ebola in the Democratic Republic of the Congo (2018-present), Pneumonic Plague in Madagascar (2017), and Cholera in Yemen (2016-2017). We propose three simple data-driven predictive models: a linear autoregression that uses historic epidemiological data to produce real-time disease activity estimates (AR), a linear regression that leverages observed Google query volumes to estimate disease incidence (GT), and a regression on both historic epidemiological data and search query data (ARGO). We find that ARGO provides useful estimates of disease activity for Yellow Fever in Angola, Zika in Colombia, and Plague in Madagascar weeks earlier than traditional healthcare-based surveillance data. We find that our data-driven methods are less effective at tracking Ebola in the DRC and Cholera in Yemen, and hypothesize that issues of sample bias and skew in search query volumes as a result of media coverage may contribute to a poor signal in these cases.

Results

Motivation for Digital Epidemiological Methods. To motivate the use of digital data streams to monitor emerging outbreaks, we produced a series of correlations assessing the relationship between each country's epidemiological curve and the volume of a simple Google search term querying the disease of interest (e.g. the search term "Zika" in the case of Colombia). As shown in Fig. 1, the search volumes appear to track the time series of cases synchronously in most countries, and we observed high correlations for Angola (r=0.84, Yellow Fever), Colombia (r=0.80, Zika), and Madagascar (r=0.73, Plague), suggesting the potential utility of digital data-driven epidemiological models.

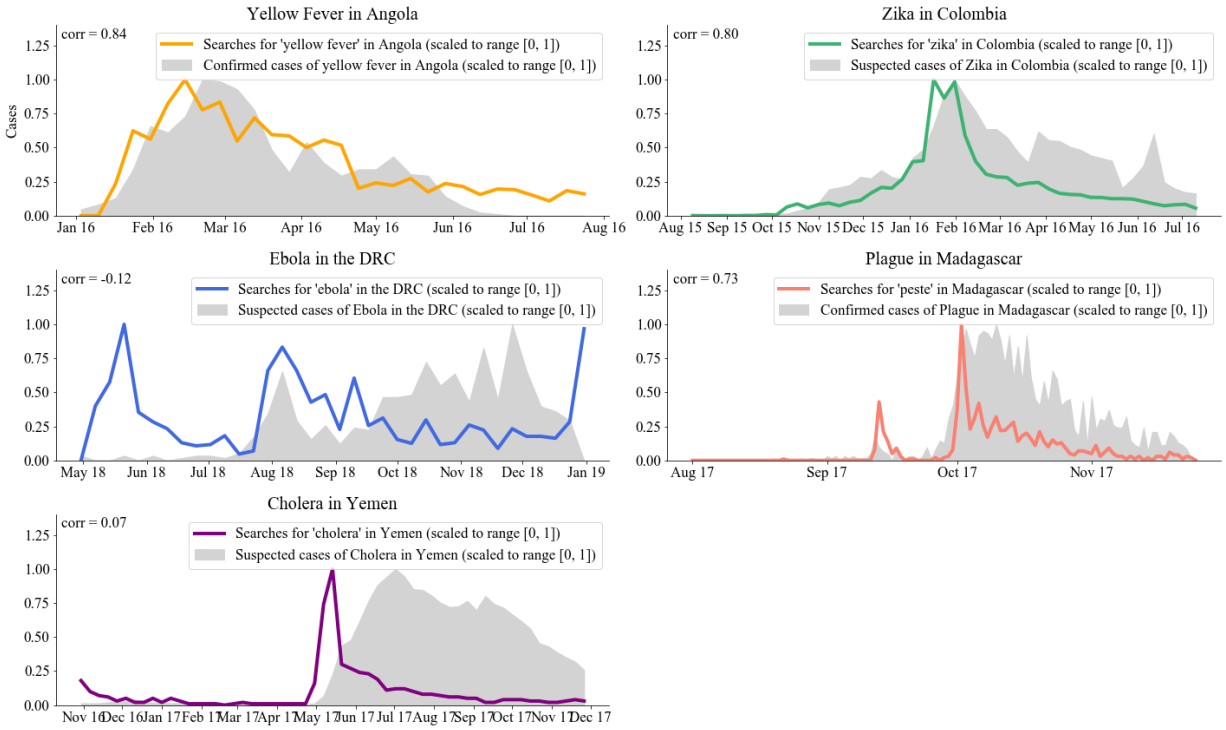

Figure 1: Motivation for digital epidemiological modeling of five emerging outbreaks. In each case, the outbreak's epidemiological curve (in grey) is compared with normalized search volumes for a single related search term in the country of question.

For each disease outbreak, we built three machine learning models to produce (retrospective and out-of-sample) real-time disease activity estimates that use input information that would have been available at the time of prediction. Our three models were trained dynamically on a continuously expanding time window to incorporate new information as it became available and are summarized as follows: (1) Autoregressive model (AR), that uses only historical cases from n weeks in the past to predict current cases; (2) Google search trends (GT), a multivariate model that uses only synchronous Google search terms for prediction; and (3) ARGO, a multivariate model similar to the one presented in [13] that combines both autoregressive case information and Google searches to make predictions. We assessed the predictive performance of each model when compared to subsequent observations by healthcare-based disease surveillance systems. Details of model implementation can be found in the Methods section.

Evaluation Assuming Continuous Flow of Available Epidemiological Data. As a reality check, our first series of models compare nowcasts 1- and 2-weeks ahead of the release of case reports with the ground truth incidence available retrospectively in weekly epidemiological updates produced by local health authorities. These models were trained and built with a strategy similar to the one used in endemic and seasonal outbreaks to make sure our efforts could produce meaningful disease estimates under the assumption that disease activity reports become available with delays of one to two weeks and are continuously available. This assumption is not always satisfied in

emerging disease outbreaks. Fig. 2 shows these predictions over the full time series of each outbreak, while Table 2 summarizes the out-of-sample predictive performance across models and countries as captured by Pearson's correlation (CORR), root mean squared error (RMSE), and relative root mean squared error (rRMSE).

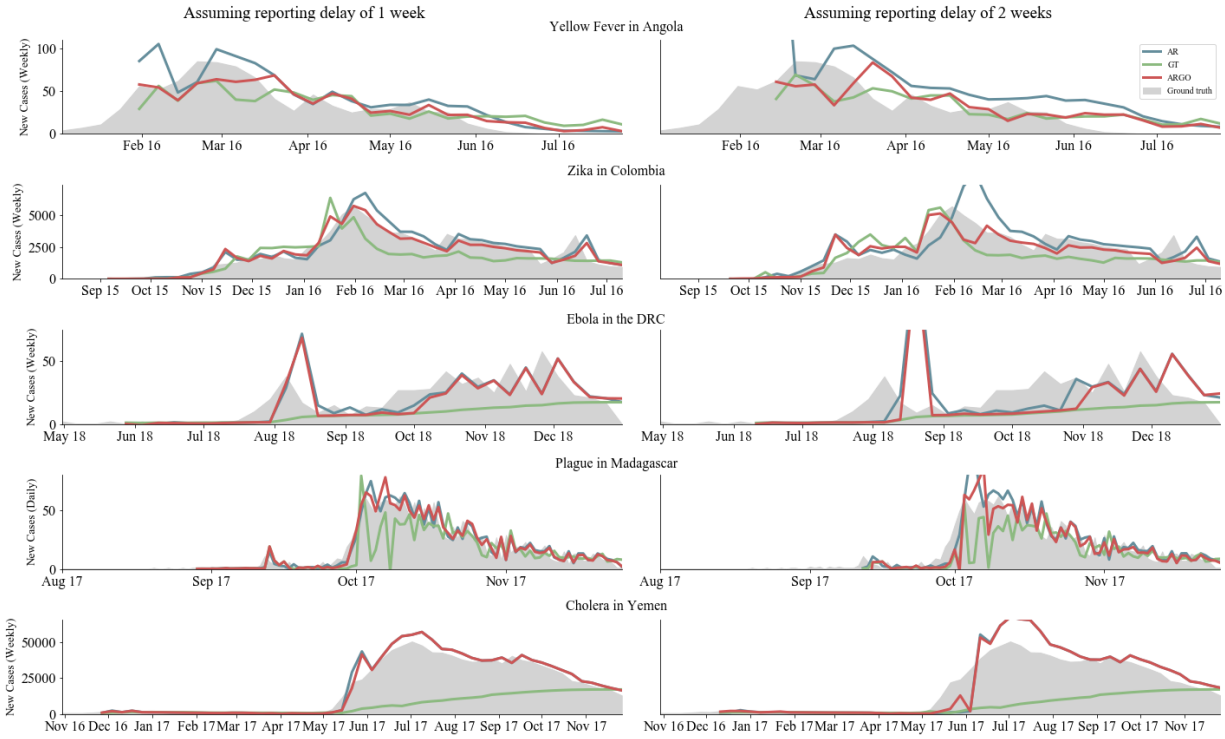

Figure 2: Series of plots comparing the nowcasts produced by three digital epidemiological models (available in real-time) to "ground truth" epidemiological data (available at a delay). The left column shows how models perform assuming a 1-week reporting delay in the traditional surveillance system; the right columns shows model performance assuming a 2-week reporting delay.

		Yellow Fever		Zika		Ebola		Plague		Cholera	
Delay (weeks)		1	2	1	2	1	2	1	2	1	2
CORR	AR	0.879	0.54	0.92	0.78	0.57	0.19	0.91	**0.88**	0.98	0.93
	GT	0.79	**0.80**	0.78	0.73	**0.582**	**0.50**	0.74	0.68	0.65	0.59
	ARGO	**0.882**	0.69	**0.93**	**0.82**	0.581	0.17	**0.92**	0.84	**0.99**	**0.94**
RMSE	AR	17.60	62.65	644.24	1176.74	15.252	28.11	8.45	**11.65**	4224.88	9156.57
	GT	17.66	**17.63**	997.45	1072.01	16.98	**18.13**	13.60	15.38	18532.22	19486.67
	ARGO	**13.22**	20.42	**542.39**	**823.34**	**15.246**	27.41	**7.97**	11.85	**3973.06**	**8497.43**
rRMSE	AR	0.55	2.10	0.31	0.54	**0.81**	1.40	0.45	**0.53**	0.23	0.48
	GT	0.56	**0.59**	0.58	0.50	0.90	**0.90**	.72	0.70	1.01	1.03
	ARGO	**0.42**	0.69	**0.26**	**0.38**	**0.81**	1.37	**0.42**	0.54	**0.22**	**0.44**

Table 1: Evaluations of three computational models (AR, GT, and ARGO) across five outbreaks, based on correlation (CORR), root mean squared error (RMSE), and relative root mean squared error (rRMSE). The result of the best-performing model for each prediction scenario and metric is bolded. It is important to note that the units of the error (RMSE) are different given that the magnitude of each outbreak was different. The relative error, however, is comparable across outbreaks.

We found that, based on RMSE and correlation, digital epidemiological models that incorporated Google information (GT and ARGO) led to reasonable disease estimates that were within range of the observed disease activity. Specifically, GT and ARGO outperformed a naïve autoregressive approach (AR) in all outbreaks and prediction horizons besides plague, in which a pure AR model performed best for 2-week delays. In general, ARGO exhibited the lowest RMSE and highest correlation in a majority of countries and prediction horizons, though Google data alone improved predictions in the case of 2-week delays in two of the outbreaks (Yellow Fever and Ebola). We note, however, that nowcast models were generally not skillful enough to track Ebola in the DRC, which exhibited substantially lower predictive performance compared to the other countries (correlation range: 0.17-0.58). Moreover, we observe that the ARGO method does not improve significantly upon a naïve autoregressive approach for tracking both Ebola in the DRC and Cholera in Yemen.

To assess the predictive power of the Google search terms used to nowcast cases each week and visualize changes in predictive power over the course of the epidemic, the size of ARGO model coefficients for each week of prediction are shown for each country in Figs. 3, S1-S5. Because the models are dynamically trained on a 1-week expanding time window, the predictive power of the variables are seen to fluctuate over the weeks of the outbreak, with many search terms appearing most important for prediction in early stages of the outbreak.

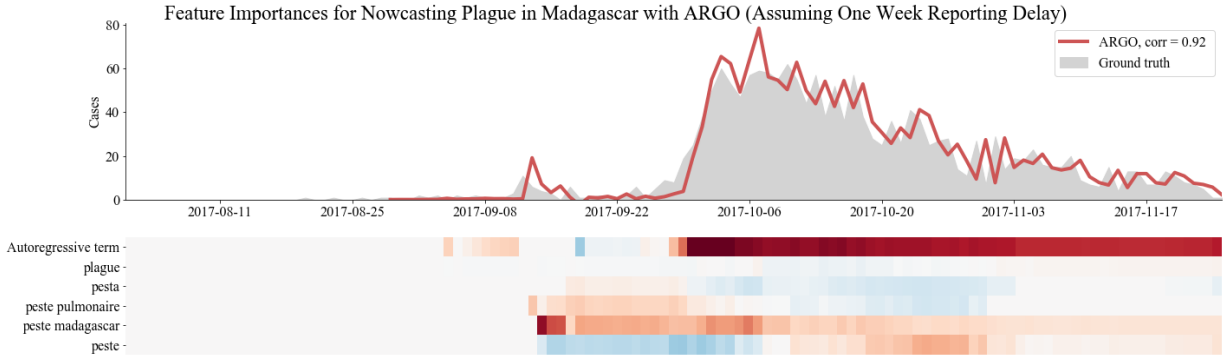

Figure 3: Evaluating feature importances (coefficients in linear regression) in ARGO for nowcasting Plague in Madagascar assuming a reporting delay of one week. Since the model is trained dynamically, feature importances shift from week to week. Note that the autoregressive term is extremely important, but information from Google search trends is also used, particularly early on in the outbreak.

Evaluation Based on Publicly Released Reports. The first evaluation approach assumes that the ground truth (weekly cases) are reported accurately within 1-2 weeks of occurrence, which is rarely the case in emerging outbreaks in which surveillance may be constrained by limited resources.

Here, we evaluate the performance of the same three models (AR, GT, and ARGO) under more realistic conditions, using partial and unrevised case reports as they were released in real-time (Fig. 3). In contrast to the first approach, here models are trained on a potentially (and frequently) unreliable ground truth, since future revisions of past disease activity may continually update case reports that are released at any given point. We assessed the feasibility of these models in achieving an estimate of disease activity when there are no epidemiological data available in real-time. This analysis was performed on all 7-20 reports for each of the five disease outbreaks; a selection of case studies are presented here and full charts are included in Figs. S6-S10.

As shown in Figure 4, we observed that, even in these realistic circumstances, ARGO produced meaningful and within-range disease activity estimates filled the temporal gap introduced by delayed availability of epidemiological reports. Moreover, when compared to the GT and AR models, ARGO appears to most closely estimate the cases that would eventually be reported throughout each outbreak.

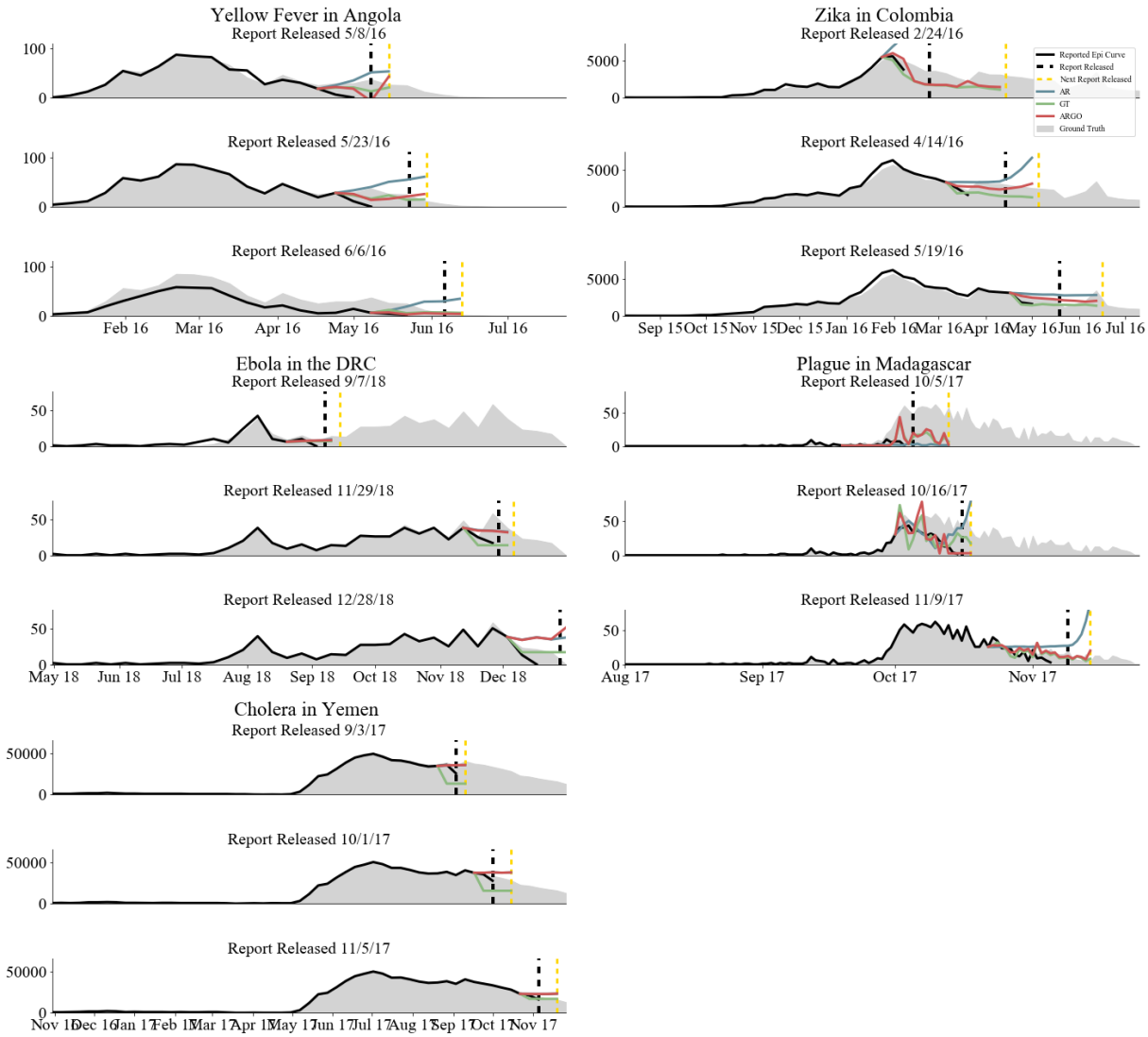

Figure 4: Summary of evaluation approach based on the historical publicly released reports. In each figure, the grey filled area is the ground truth data (available at, or after, the end of the outbreak). The black line shows surveillance data released in the report at the time of publication (the date of publication is denoted by the dashed black line), and the colored lines show the real-time predictions of our three models. Here, figures are included for three epidemiological situation reports for each outbreak; more plots with the same evaluation task can be found in Figs. S9-12.

Discussion

We have shown that machine learning techniques that combine real-time disease-related Google search activity with (delayed and frequently incomplete) epidemiological information available during emerging outbreaks can provide useful real-time insights on the likely trajectory of disease transmission. By assessing model predictions in (i) a setting that assumes the continuous availability of delayed epidemiological information (reporting delays of 1-2 weeks with no case revision) and (ii) a set of realistic historical settings where delayed information was unavailable or unreliable (reporting delays of variable week lengths and with case revisions in subsequent epidemiological reports), we demonstrate that incorporating disease-related Google search information improves predictions across several disparate disease and country contexts.

In particular, we demonstrate, for the first time, how a digital nowcast model like ARGO would be deployed in real-time during multiple distinct emerging disease outbreaks with reporting delays and surveillance revisions. We show specifically the insights that would have been accessible in real-time should our approaches have been implemented during the emergence of these outbreaks. Consider, for example, the real-time disease predictions for the 2017 plague outbreak in Madagascar shown on the right-middle panel in Fig. 4. The black line, which indicates the number of known reported cases at the time of release of an epidemiological report (Oct. 16, 2017), suggests a sharp decline in cases in October. By the end of the outbreak, it would become clear that there was no decrease in cases in October (ground truth cases produced at the end of the outbreak are shown in gray shading), an insight which was not available in real-time, but which was captured by the Google-based model (GT, green line). At the very least, our predictions have the potential to signal to health officials when outbreaks are not yet over, when cases may be increasing, and when cases may be decreasing, supporting key decision-making on large-scale treatment and prevention measures. We find that the pattern demonstrated in Madagascar generalizes to other diseases and regions: in all five of the diseases we analyze, we find that epidemiological data can be effectively supplemented in real-time with digital epidemiological methods.

In addition to showing the potential utility of real-time predictions trained on unreliable or incomplete epidemiological data, our analysis confirms the findings of other digital epidemiology studies that demonstrate the added value of combining Google-based predictions with autoregressive case information [13, 18, 16, 28]. Indeed, the ARGO coefficient heatmaps in Figs. 3, S1-S4 reveal that the epidemiological case information from previous weeks has consistently strong predictive power over the course of the outbreak, while the importance of Google predictors fluctuates over time and appears to be most useful in the earlier stages of the studied outbreaks. The phenomenon that past cases are intrinsically linked to future cases is a common feature of infectious disease outbreaks: here, we leverage this fact to improve the accuracy of our predictions, evidenced by the fact that ARGO generally outperforms the Google-only and autoregressive models across diseases and prediction horizons. Further, our findings suggest that the relative feature importance of autoregressive information and GT data is dependent on the timescale of disease transmission (serial interval). Specifically, we find that GT data appears to posses greater predictive power in diseases with short serial intervals like influenza, and less predictive power in diseases like Cholera, where transmission time-scales are typically longer.

While there are many promises of using Google data to track and predict outbreaks, there are several limitations to using Google data for epidemiological purposes. In the context of emerging outbreaks, these include bias in the sample of Google users and the bias introduced as a results of media coverage. Google users are a non-random sub-sample of the population, and this bias is particularly significant in the context of most emerging outbreaks, which occur in developing regions where Internet penetration is relatively low and in which there are significant rich-poor and urban-rural divides in Internet access. As a result, it is possible that much of the disease-related Google search activity may occur in a country's capital, while cases of the disease may occur all over the country or in a specific region with low internet penetration. Similarly, which search terms are and are not selected could bias affect performance. Exploration of Google search activity on sub-national levels could help provide insight into this issue. This bias will likely become less relevant as global internet penetration in rural regions increases.

Additionally, media coverage may confound the interpretation of our models. In using Google query volumes as a proxy for disease activity, it may be the case that queries come from individuals who are infected or suspect infection. However, we inevitably also receive signals resulting from high media coverage (often pervasive during novel and unexpected outbreaks), which prompts large numbers of people in the affected country to search for disease-related terms out of curiosity, seeking news articles. Consider the graph of search volumes for the term "peste" (French for "plague") in Madagascar in Fig. 1: there is a sharp spike in volumes in mid-October, which appears anomalous to the incidence curve. It is very reasonable to hypothesize that this spike is the result of the first media coverage of that outbreak.

To evaluate how media coverage may skew Google search volumes, we qualitatively compare signals in Google searches and news report volumes with epidemiological time series. Figure 5 compares the volume of news articles (obtained from the GDELT Global Knowledge Graph [29]), Google search trends, and reported cases side by side for each outbreak. Based on this analysis, it is plausible that ARGO's weaker performance on Ebola in the DRC and on Cholera in Yemen are caused by premature spikes in Google searches. These premature spikes are correlated with early spikes in news coverage, and these early spikes are not found for the other outbreaks where ARGO had better performance. It is likely that hype caused by media coverage biases predictions based on Google search volumes in these analyses.

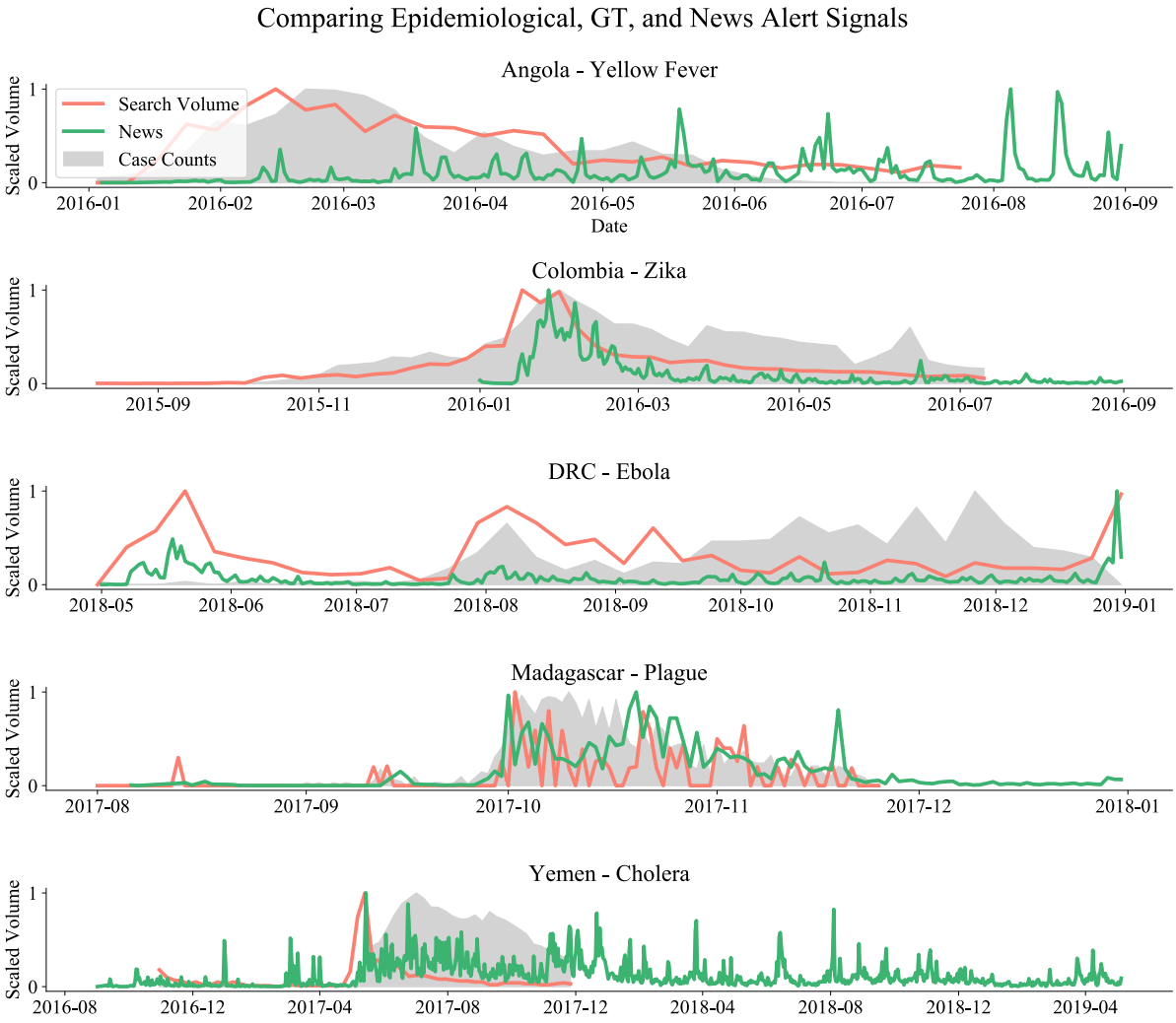

Figure 5: Comparison of signals in ground-truth epidemiological data, Google search query volumes, and news alerts data from the GDELT Global Knowledge Graph. Note how media coverage (as captured in the news alerts time-series) may bias predictions based on the GT data.

Here we have shown how Internet-based data streams can be mined to monitor the progression of emerging outbreaks in low-income settings where traditional surveillance may lag substantially or be rendered inaccurate due to backfilling. We have shown that digital epidemiological methods like ARGO perform well for nowcasting plague in Madagascar, Yellow Fever in Angola, and Zika in Colombia, but are less effective at tracking Cholera in Yemen and Ebola in the DRC. The poor performance for the Ebola and Cholera outbreaks could be linked to a combination of low internet coverage, intense response to news alerts, and rapid shifts in disease dynamics due to population unrest and violence. Future work should focus on the pathogen and population conditions (digital coverage, symptoms specificity, serial interval, mode of transmission, behavior changes, and health

interventions) that can make or break digital surveillance in low-income settings, and how to adjust digital surveillance signals for intense media coverage and other exogenous forces.

Materials and Methods

Data Sources

We digitized daily or weekly national case counts from epidemiological situation reports for outbreaks of Yellow Fever in Angola (Jan. 3 - July 31, 2016), Zika in Colombia (Aug. 9, 2015 - July 10, 2016), Ebola in the Democratic Republic of the Congo (April 30 - Dec. 31, 2018), Pneumonic Plague in Madagascar (Aug. 1 - Nov. 2016), and Cholera in Yemen (Oct. 30, 2016 - Nov. 26, 2017). We also downloaded country-specific time-series of Google query volumes from the Google Trends API for the same time periods.

Epidemiological Data — The following table summarizes the sources of epidemiological data and key descriptive statistics on the epidemiological dataset for each of the five outbreaks analyzed. For each dataset, we consider the final epidemiological report to be the "ground truth" recording the true onset date for each of the cases in the outbreak; the earlier reports are considered estimates and subject to revision. Note that this assumption requires a larger leap for Ebola and Cholera than for the other outbreaks analyzed, as these outbreaks were ongoing at the time of data collection whereas the other outbreaks were completed. Finally, note that, due to issues of data availability, in certain outbreaks the dataset consists of only laboratory-confirmed cases, while in other outbreaks the dataset contains both confirmed and probable (or suspected) cases.

Table 2: Epidemiological Data Sources

Outbreak	Time Period	Temporal Granularity	Total Cases	Reports	Source
Yellow Fever in Angola	Jan. 3 - July 31, 2016	Weekly	879 (confirmed)	11	Digitized from plots in PDF situation reports released by WHO [5]
Zika in Colombia	Aug. 9, 2015 - July 10, 2016	Weekly	91,156 (suspected)	7	Digitized from plots in PDF epidemiological updates published after Feb. 17 (only updates with Colombia-specific data are included) [6]
Ebola in the DRC	Apr. 30 - Dec. 31, 2018	Weekly	628 (suspected)	17	Digitized from plots in PDF situation reports released by the WHO [7]
Pneumonic Plague in Madagascar	Aug. 1 - Nov. 25, 2016	Daily	1,857 (confirmed)	12	Digitized from plots in PDF situation reports released by the IPM [9] and WHO [8] (only reports containing case counts specifically for Pneumonic Plague are included)
Cholera in Yemen	Oct. 30, 2016 - Nov. 26, 2017	Weekly	973,802 (suspected)	13	Digitized from plots in PDF situation reports released by WHO AFRO [10]

Google Search Trends Data — Time-series downloaded from Google search trends [30] describe the number of people searching for a specific keyword, in a specified geographic region, each day, week, or month (normalized to a 0 - 100 range). Google search trends data was extracted for each outbreak for the same time period as the epidemiological data, on the same temporal granularity as the epidemiological data, and limited to searches in the country of the outbreak. To avoid forward-looking bias, it is standard to select keywords by using Google correlate to find search terms that correlate well with the epidemiological time-series in a training period (which is then not included in the evaluation period) [13, 18, 28]. However, since Google correlate data is not available for any of the countries we analyze, we select a few simple keywords for each outbreak that are clearly related to the disease in question. In certain cases, there is not enough Google search information to yield meaningful results in the sample available through Google search trends: for example, we identified "fièvre hémorragique" and "fievre hemorragique" as relevant search terms for Ebola in the DRC, but were unable to include them due to a lack of available search signal. Similarly, we experimented with including "diarrhea" and the Arabic versions of "cholera" and "diarrhea" for the outbreak of Cholera in Yemen, but did not find an improvement in signal over using only "cholera" in English.

Table 3: Search Terms by Outbreak

Outbreak	Search Terms
Yellow Fever in Angola	'yellow fever', 'febre amarela'
Zika in Colombia	'zika', 'zika sintomas', 'el zika', 'sintomas del zika', 'virus zika', 'zika colombia', 'el zika sintomas', 'el sica'
Ebola in the DRC	'ebola'
Plague in Madagascar	'plague', 'pesta', 'peste', 'peste pulmonaire', 'peste madagascar'
Cholera in Yemen	'cholera'

News Alert Data — News alert data was obtained from the GDELT Global Knowledge Graph in the form of fractions of daily raw article counts that are relevant to a query. GDELT is a large and regularly updated open database and platform that monitors the world's news media in over 100 languages [29].

Models

We explored three simple data-driven nowcasting models, emphasizing model simplicity as there is often not enough data available in emerging outbreaks to train a more complex model.

Linear Autoregression (AR) — An autoregressive model uses a linear combination of past observations of disease incidence ("autoregressive terms") to provide an estimate for synchronous incidence. Here, we choose for simplicity to use only the single most recently observed autoregressive term, so the linear autoregression is a univariate linear regression:

$$y_t = \beta y_{t-h} + \alpha \tag{1}$$

The linear regression is optimized over available training observations to minimize mean squared error loss. The time horizon of prediction h depends on the reporting delay in each outbreak; for instance, if there is a two-week reporting delay in a surveillance system, the autoregressive term

will be the 2-week lag, so $h = 2$.

Regression on Google Query Volumes (GT) — Our second model is a multivariate regression mapping synchronous data on Google query volumes for selected search terms to estimated synchronous incidence. Depending on the number of search terms selected for each outbreak, this regression contains 1-8 variables.

$$y_t = \Sigma_{g \in G} \beta_g g + \alpha \tag{2}$$

We adopt a L1 regularization to prevent overfitting and provide automatic feature selection, with the regularization parameter selected via 5-fold cross validation on the training set from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The LASSO regression is optimized over available training observations to minimize mean squared error loss.

Autoregression and Regression on Google Query Volumes (ARGO) — ARGO combines the AR and GT methods in a single multivariate regression including both a single autoregressive term (the most recently observed incidence value) and a set of synchronous Google query volumes.

$$y_t = \beta y_{t-h} + \Sigma_{g \in G} \beta_g g + \alpha \tag{3}$$

As in GT, ARGO is made more robust with L1 regularization, with the regularization parameter selected via 5-fold cross validation on the training set from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The ARGO method used here is a somewhat simplified version of the linear regression on autoregressive data and synchronous Google query data originally developed to nowcast influenza in the United States [13].

Evaluation

We had access only to publicly released epidemiological situation reports, which are typically released somewhat sporadically, exhibiting long reporting delays and gaps where no information is available at all. To capture two possible data-access scenarios, (1) an ideal scenario in which final case numbers are reported 1-2 weeks after they occur, and (2) a more realistic scenario in which case numbers are reported with some delay and possibly corrected at a later date, we adopted two separate methods of evaluation. The first evaluation method assumes a continuous flow of correct epidemiological data and a set reporting delay of one to two weeks. The second method reflects the reality of many epidemiological reporting systems by using the data presented in publicly released epidemiological reports.

Evaluation Assuming Continuous Flow of Epidemiological Data — The first form of evaluation uses only a single time-series of epidemiological data; the "ground truth" (taken as the last epidemiological report on the outbreak publicly released). We assumed a h-week reporting delay and experiment with h taking on values of 1 and 2. Thus this evaluation method represents a near-ideal data access scenario in which case counts, once reported, are never adjusted or corrected. We adopted dynamic training (also known as online learning or walk-forward validation) so that, when predicting each week's incidence, each of the models is trained on all the data available up to that week. Models were then evaluated over the entire time-series based on Pearson's Correlation Coefficient (CORR), root mean squared error (RMSE), and relative root mean squared error

(rRMSE).

$$CORR = \frac{\Sigma_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - x_i)^2} \tag{5}$$

$$rRMSE = \frac{\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - x_i)^2}}{\bar{y}} \tag{6}$$

Evaluation Based on Publicly Released Epidemiological Situation Reports — The ideal data-access scenario described above is not always the case in emerging outbreaks, which are characterized by reporting gaps and revisions of case counts after initial publication. The second method of evaluation recognizes this challenge, and compares the accuracy and timeliness of epidemiological reports that were publicized in each outbreak with the accuracy and timeliness of our three digital epidemiological models. We first empirically estimated the average reporting delay for each outbreak as the average number of days or weeks from initial reporting to a stable count of cases for a given day or week of the outbreak in the epidemiological reports. To account for small human errors in reporting and digitization of reports, we defined a "stable" case count as one that does not change by more than 1% from one week to the next. In practice, we observed a 2-week reporting delay for all five outbreaks presented. Note that while this empirical method requires several weeks of published epidemiological reports, a healthcare system's reporting delay could likely be estimated a priori by its managers.

For each report released during each outbreak, we trained the three listed digital epidemiological models on the data that was stable in the report (according to the calculated reporting delay). We trained models for every time horizon between when stable data in the report ceased to be available and when the next epidemiological report was posted (as a way to evaluate what utility digital epidemiological models would have had at the time). Since in much of this period there was no ground truth data available, there is no simple way to evaluate the quality of our models in comparison to traditional surveillance methods for this evaluation scenario. However, we present the graphs of this evaluation method for qualitative analysis.

Tools and Code Availability

All models and evaluation metrics are implemented in Python 3.6 with scikit-learn 0.19.1. All scripts and data used in this study are publicly available at `https://github.com/emilylaiken/outbreak-nowcasting`.

Acknowledgements

This study was funded in part by the Bill and Melinda Gates Foundation (OPP 1195154). This study does not necessarily represent the views of the NIH or the US government.

Supporting Materials

Analyses of Feature Importance

Figure S1: Feature importance heatmaps for nowcasting Yellow Fever in Angola with ARGO.

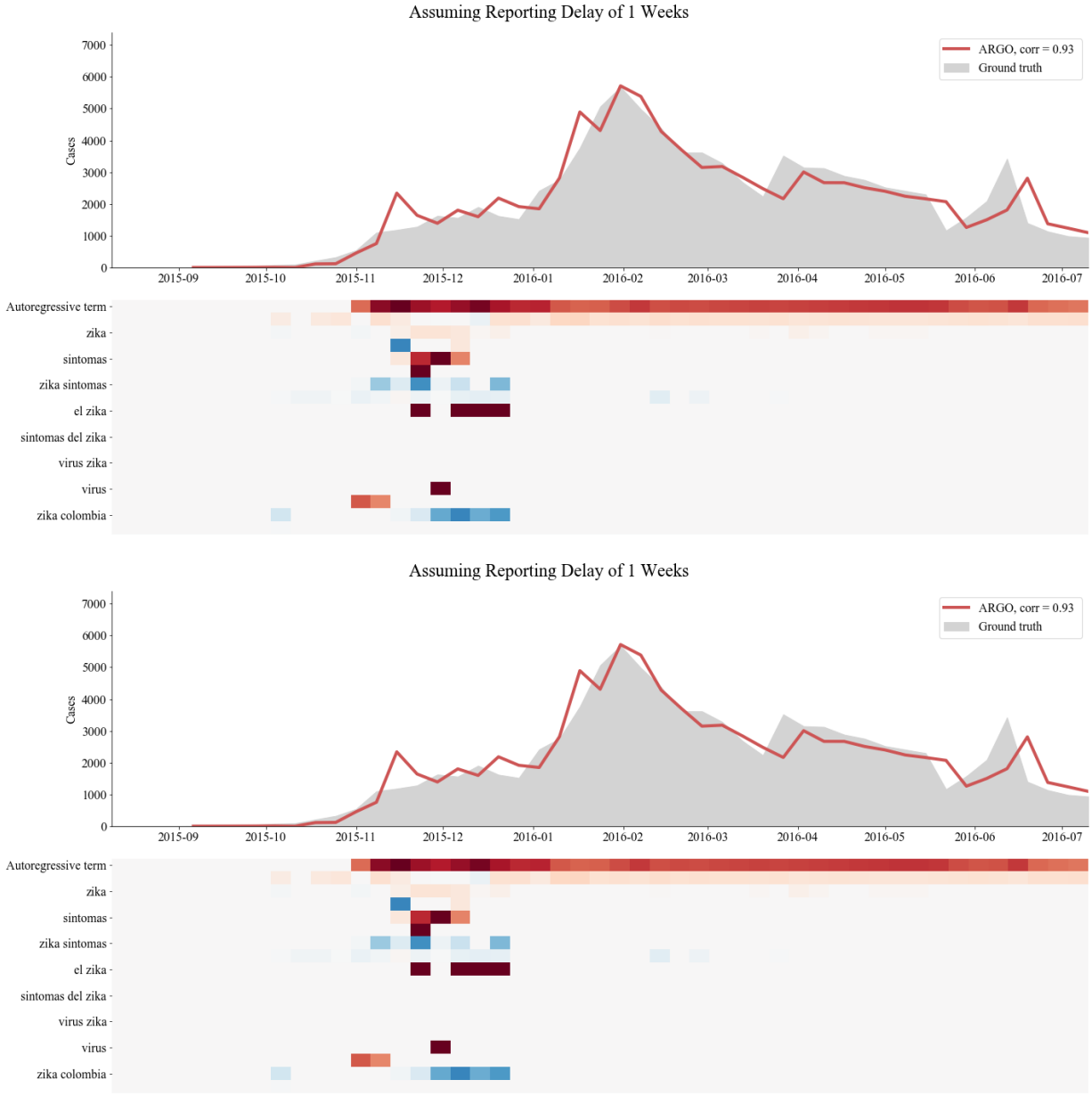

Figure S2: Feature importance heatmaps for nowcasting Zika in Colombia with ARGO.

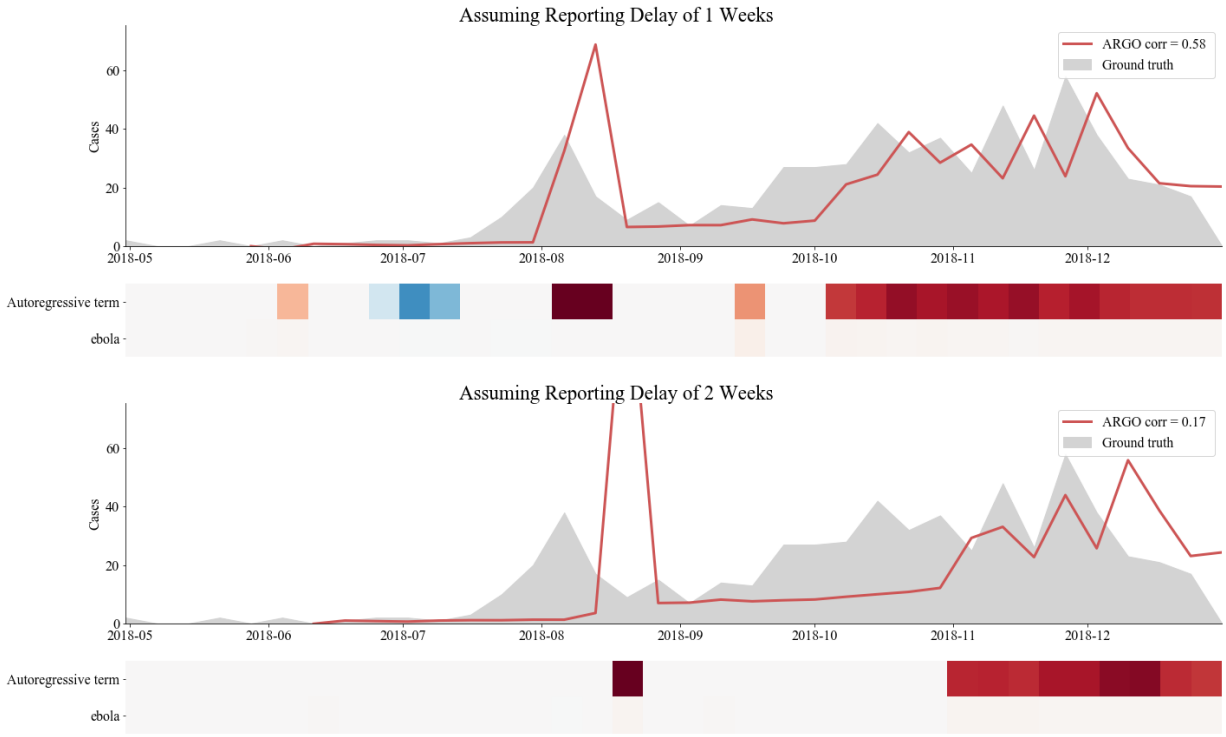

Figure S3: Feature importance heatmaps for nowcasting Ebola in the DRC with ARGO.

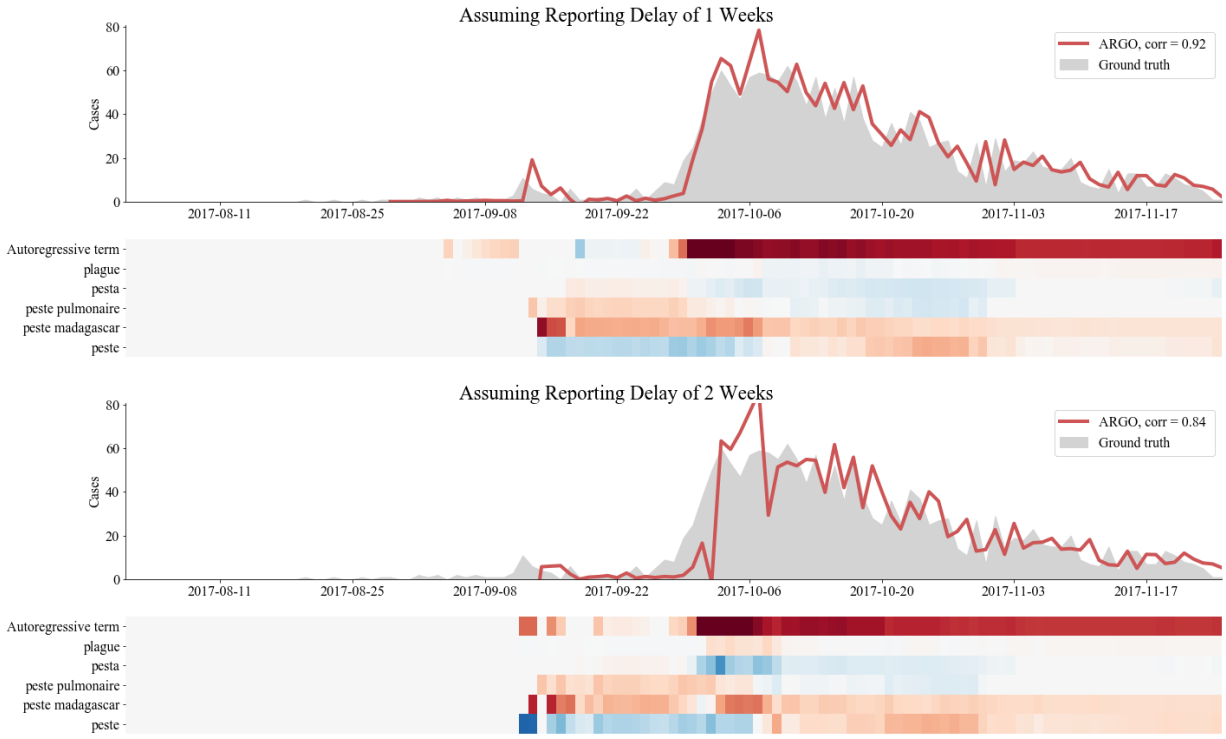

Figure S4: Feature importance heatmaps for nowcasting Plague in Madagascar with ARGO.

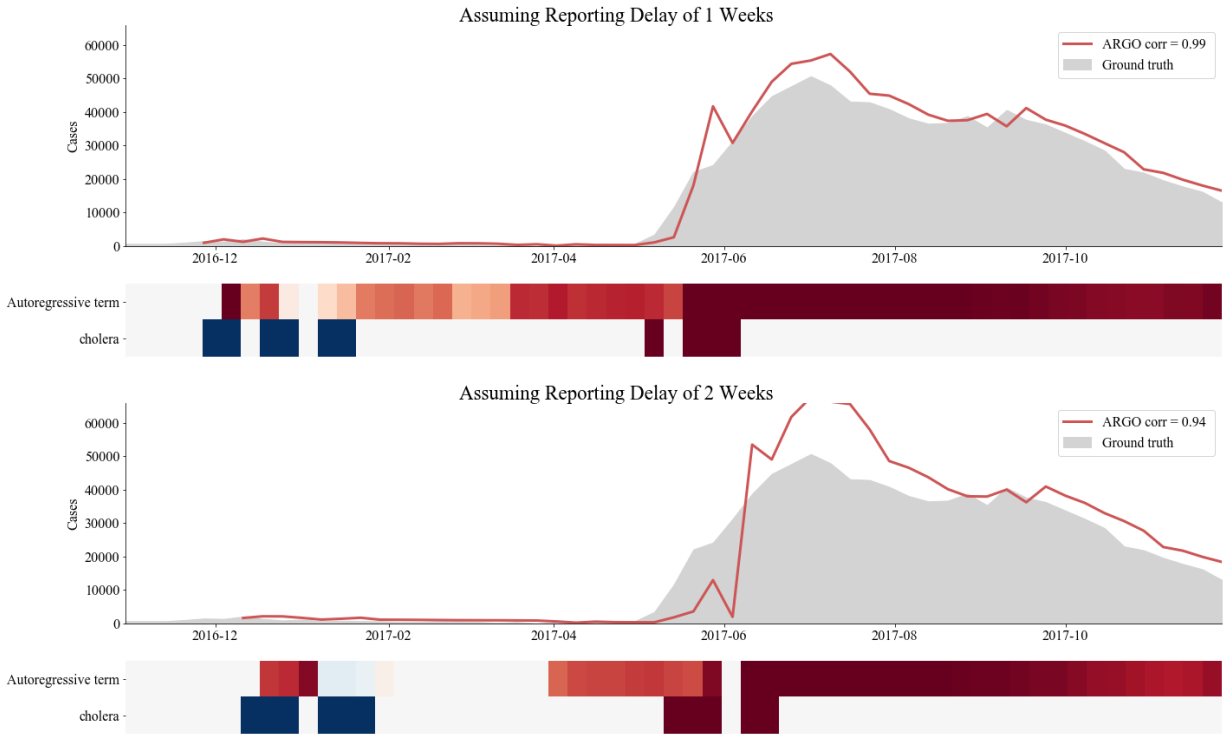

Figure S5: Feature importance heatmaps for nowcasting Cholera in Yemen with ARGO.

Additional Comparisons of Epidemiological Situation Reports and Digital Epidemiological Models

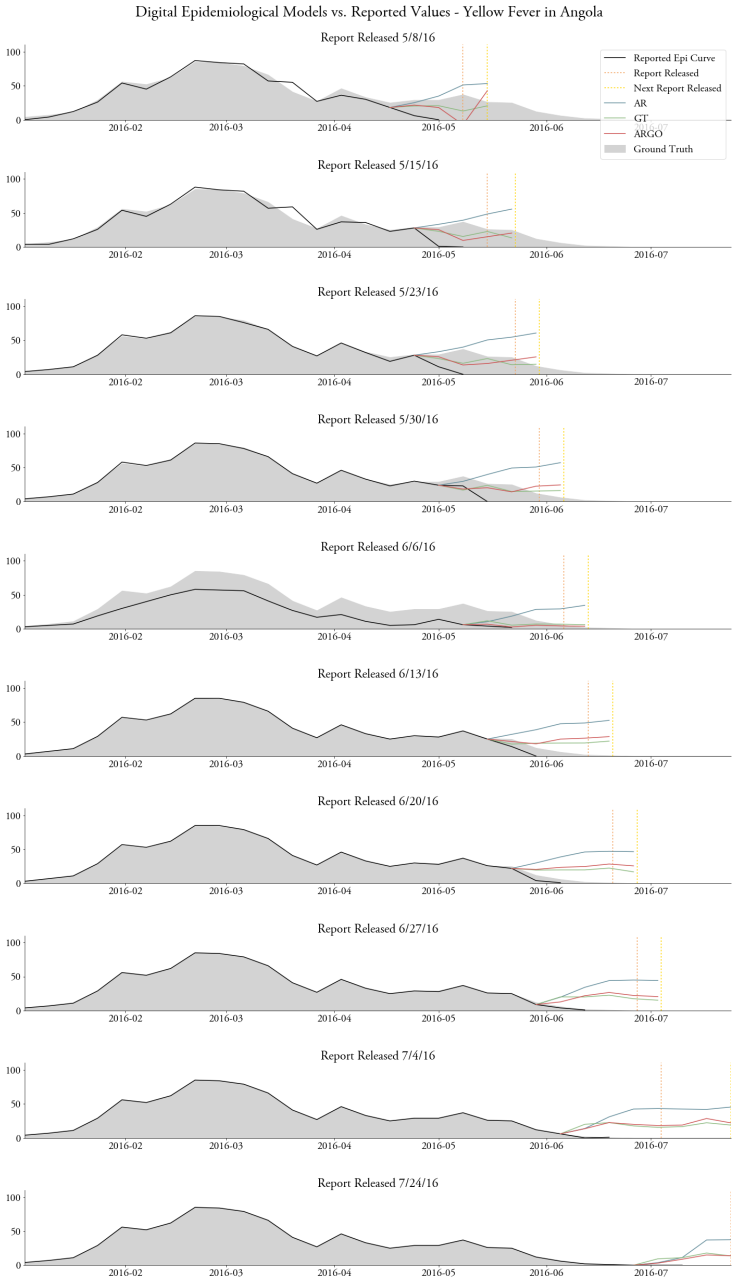

Figure S6: Comparing the accuracy and timeliness of publicly released epidemiological updates from the outbreak of Yellow Fever in Angola to the accuracy and timeliness of our digital epidemiological models.

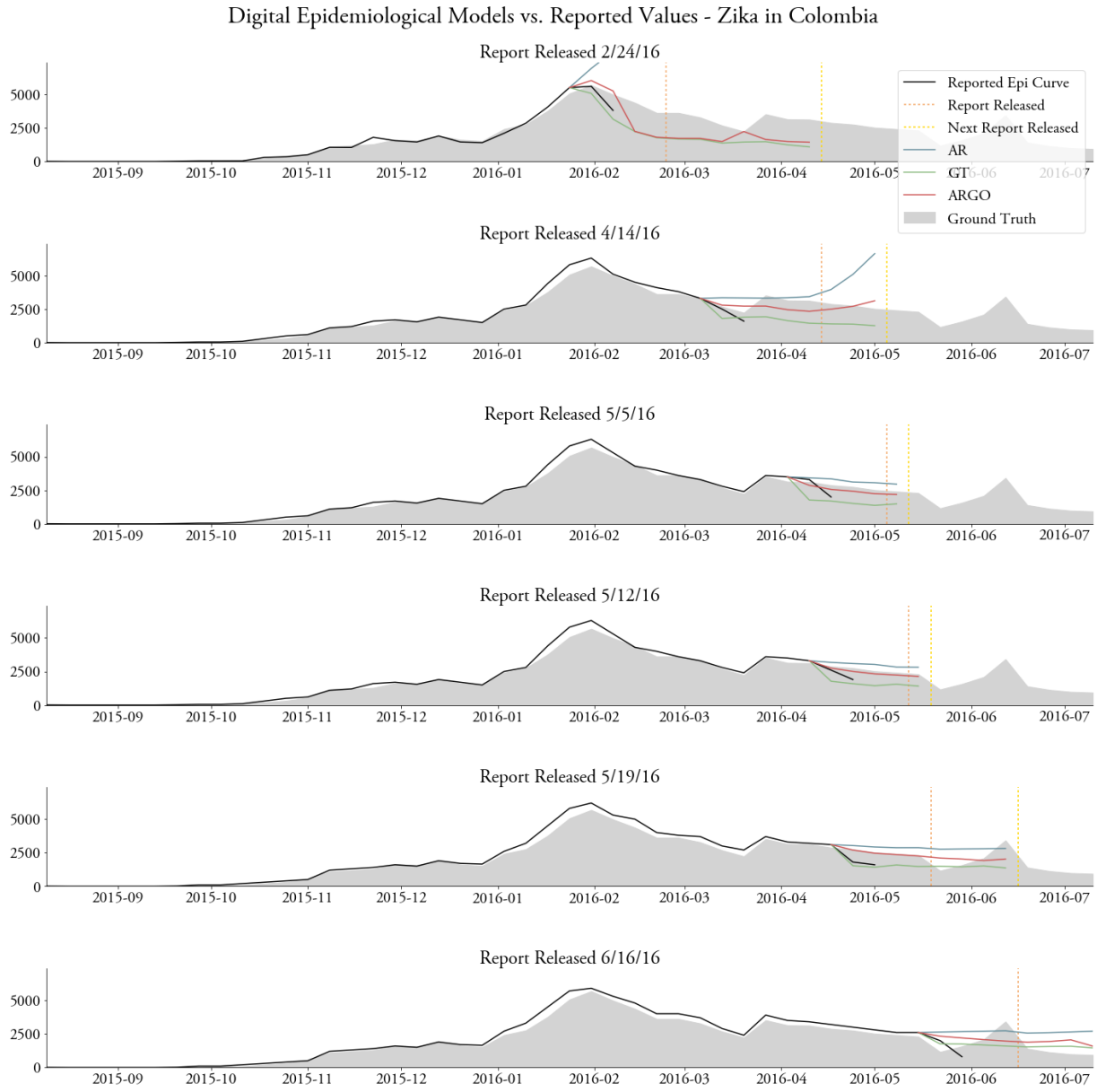

Figure S7: Comparing the accuracy and timeliness of publicly released epidemiological updates from the outbreak of Zika in Colombia to the accuracy and timeliness of our digital epidemiological models.

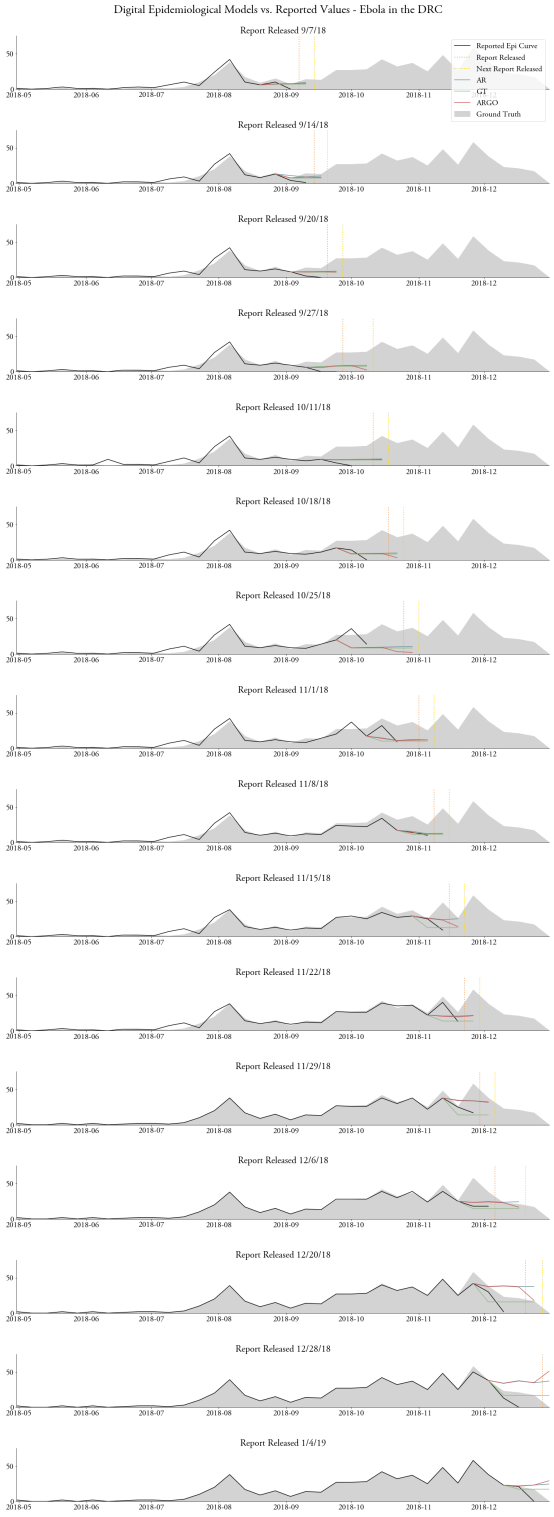

Figure S8: Comparing the accuracy and timeliness of publicly released epidemiological updates from the outbreak of Ebola in the DRC to the accuracy and timeliness of our digital epidemiological models.

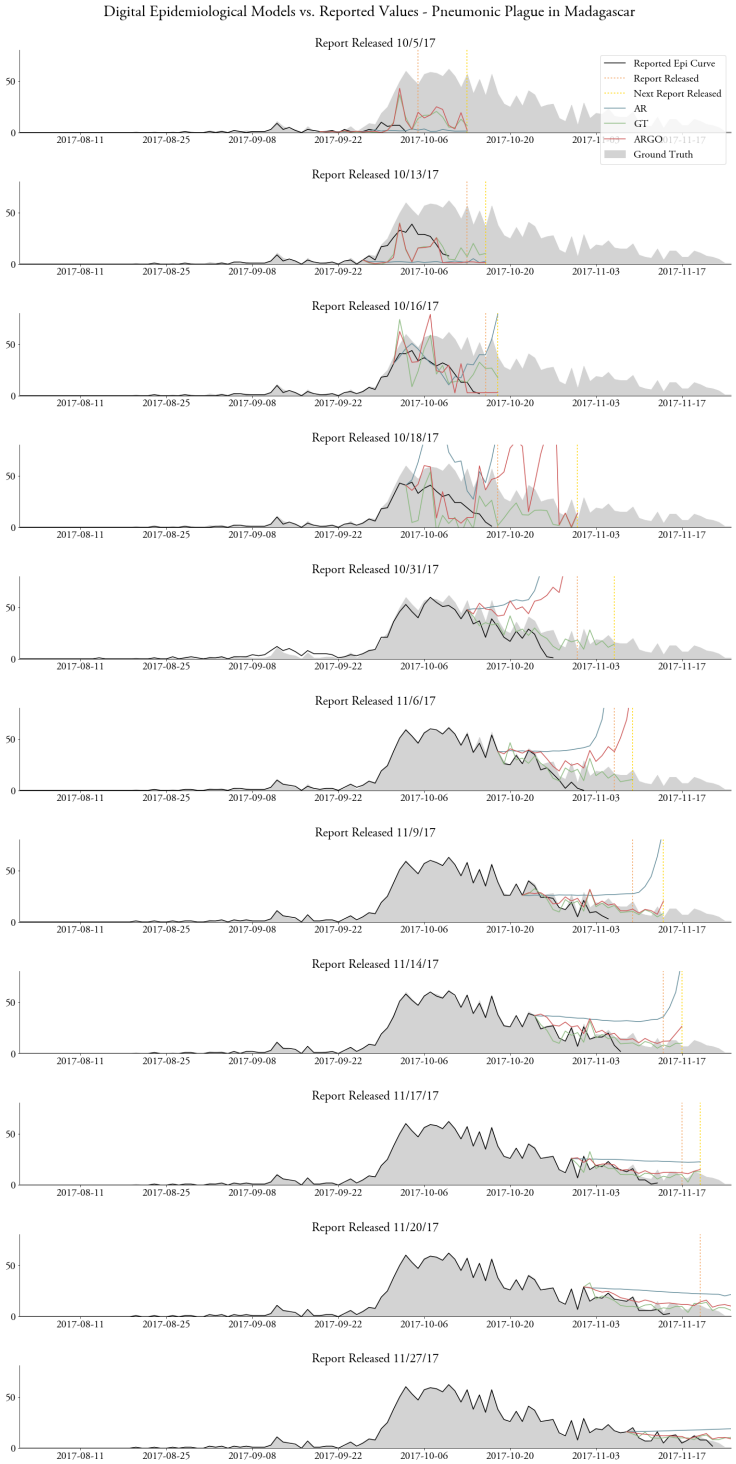

Figure S9: Comparing the accuracy and timeliness of publicly released epidemiological updates from the outbreak of Plague in Madagascar to the accuracy and timeliness of our digital epidemiological models.

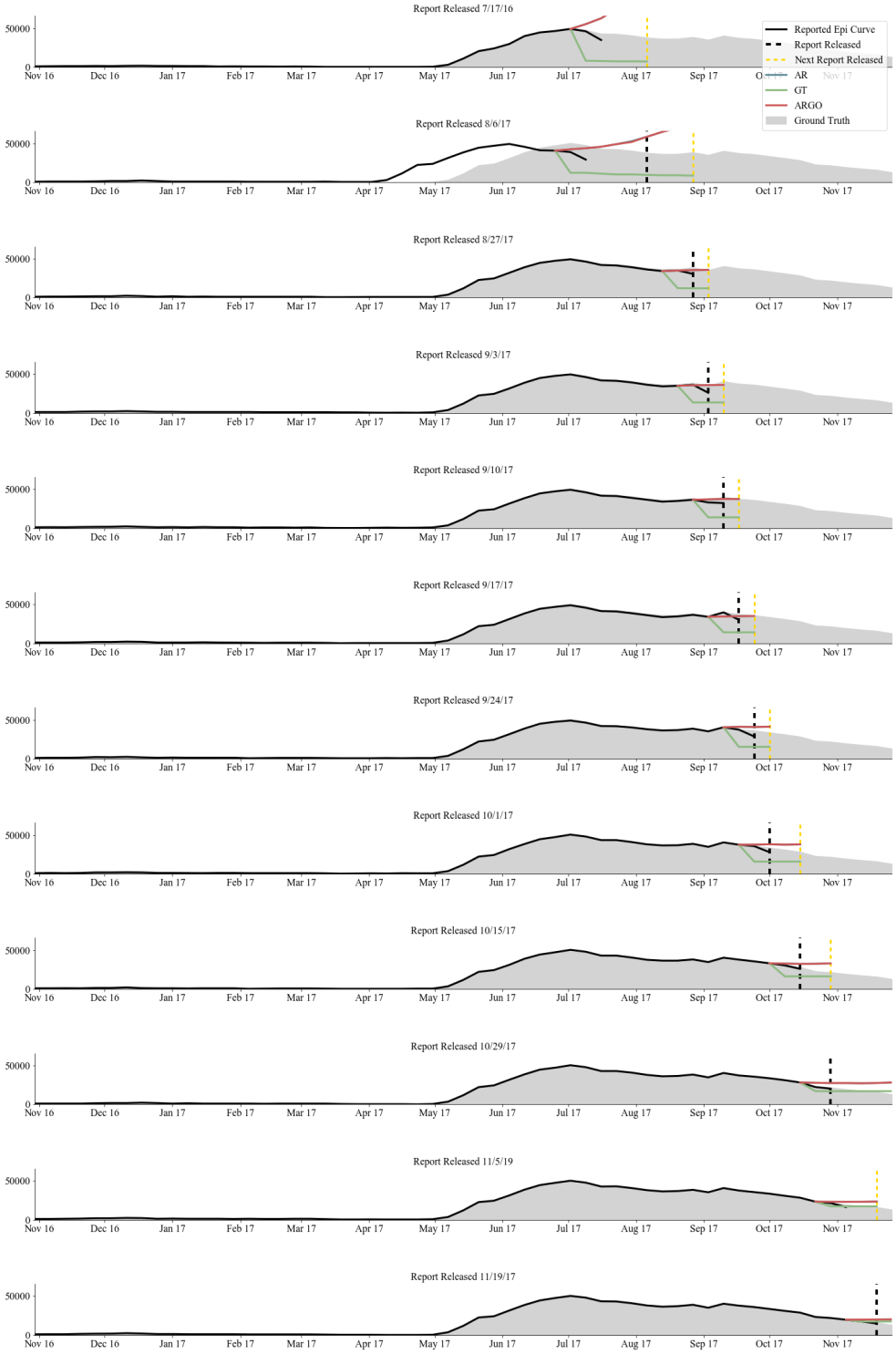

Figure S10: Comparing the accuracy and timeliness of publicly released epidemiological updates from the outbreak of Cholera in Yemen to the accuracy and timeliness of our digital epidemiological models.

References

[1] Lipsitch M & Santillana M. Enhancing Situational Awareness to Prevent Infectious Disease Outbreaks from Becoming Catastrophic. In: Inglesby T Global Catastrophic Biological Risk. *Current Topics in Microbiology and Immunology*. Springer, Berlin, Heidelberg (2019).

[2] Lipsitch M et al. "Improving the Evidence Base for Decision Making During a Pandemic: The Example of 2009 Influenza A/H1N." *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 9 (2011).

[3] Probert W et al. "Real-time decision making during emergency disease outbreaks." *PLOS Computational Biology* 14, e1006202 (2018).

[4] Brooks L et al. Flexible Modeling of Epidemics with an Empirical Bayes Framework. *PLOS Computational Biology* 11, 1004382 (2015).

[5] World Health Organization. "Yellow fever situation reports," https://www.who.int/emergencies/yellow-fever/situation-reports/archive/en/.

[6] Pan American Health Organization. "Archive by Disease - Zika virus infection," https://www.paho.org/hq/index.php?option=com_content&view=article&id=10898:2015-archive-by-disease-zika-virus-infection

[7] World Health Organization. "Ebola situation reports: Democratic Republic of the Congo," https://www.who.int/ebola/situation-reports/drc-2018/en/.

[8] World Health Organization Regional Office for Africa. "Plague outbreak situation reports," https://www.afro.who.int/health-topics/plague/plague-outbreak-situation-reports.

[9] Institut Pasteur de Madagascar. "Synthese des résultats biologiques Peste," http://www.pasteur.mg/wp-content/uploads/2017/11/20171114_Bulletin_Peste_IPM_14112017_V5.pdf.

[10] World Health Organization Regional Office for the Eastern Mediterranean. "Cholera," http://www.emro.who.int/pandemic-epidemic-diseases/cholera/index.html.

[11] Majumder, M & Rose, S. "Vaccine Deployment and Ebola Transmission Dynamics Estimation in Eastern DR Congo" (2018). Available at SSRN: https://ssrn.com/abstract=3291591.

[12] Ginsberg J et al. Detecting influenza epidemics using search engine query data. *Nature* 457, 1012-1014 (2009).

[13] Yang S, Santillana M, & Kou SC. Accurate estimation of influenza epidemics using google search data via ARGO. *Proceedings of the National Academy of Sciences* 112, 14473-14478 (2015).

[14] Santillana M et al. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology* 11, e1004513 (2015).

[15] Lu F et al. Accurate influenza monitoring and forecasting in the Boston metropolis using novel Internet data streams. *Journal of Medical Internet Research* 4, e4.7 (2018).

[16] Lu F et al. Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. *Nature Communications* 10, 147 (2019).

[17] Clemente LC, Lu F, & Santillana M. Improved real-time influenza surveillance using Internet search data in eight Latin American countries. *JMIR Public Health Surveillance* 5(2) (2019).

[18] Yang S et al. Advances in the use of Google searches to track dengue in Mexico, Brazil, Thailand, Singapore and Taiwan. *PLOS Computational Biology* 13, e1005607 (2017).

[19] Paul MJ, Dredze M, & Broniatowski D. Twitter Improves Influenza Forecasting. *PLOS Currents Outbreaks* (Oct. 28, 2014).

[20] Freifeld CC et al. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J Am Med Inform Assoc*, 15(2):150–157 (2008)

[21] Viboud C et al. Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the US. *PLoS One* 29;9(7):e102429 (2014).

[22] Santillana M et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. *Scientific reports*, 6 (2016).

[23] Smolinksi M et al. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. *American Journal of Public Health* 105, 2124-230 (2015).

[24] Paolotti D et al. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. *Clin Microbiol Infect*. 20(1):17-21 (2014).

[25] Yang S et al. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC infectious diseases*, 17(1), p.332 (2017).

[26] Majumder M et al. Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak. *JMIR Public Health Surveillance* 2 (2016).

[27] Chunara R, Andrews JR, & Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *American Journal of Tropical Medicine and Hygiene* 86, 39-45 (2012).

[28] McGough SF, Brownstein JS, Hawkins J, & Santillana M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLoS Neglected Tropical Diseases* 11, e0005295 (2017).

[29] "GDELT," `https://www.gdeltproject.org/`.

[30] "Google Trends," `https://trends.google.com/`.