Reconstructing the MERS Disease Outbreak from News

Ananth Balashankar
ananth@nyu.edu
New York University
New York, USA

Aashish Dugar
ad4025@nyu.edu
New York University
New York, USA

Lakshminarayanan Subramanian
lakshmi@nyu.edu
New York University
New York, USA

Samuel Fraiberger
sfraiberger@worldbank.org
World Bank
Washington DC, USA

ABSTRACT

Disease surveillance is critical for mobilizing health care resources and deciding on isolation measures to contain the spread of infectious diseases. Because ground truth signals of rare and deadly diseases are sparse, it can be useful to enrich surveillance systems using measures of social and environmental factors which are known to influence the spread of a disease. One approach to measure such factors is by using real time news streams. In this study, we model the epidemiological transmission of the Middle Eastern Respiratory Syndrome (MERS) disease during the outbreak that occurred from 2013 to 2018 in the Arabian peninsula. Using the GDELT news event database, we show that conflict related signals allow us to reconstruct the time series of newly infected cases per week. This reduces the residual sum of squared errors by a factor of 3.36 as compared to a standard epidemiological model. We also capture interpretable time-sensitive factors which illustrate the importance of using real time news stream to model the evolution of a disease such as MERS and facilitate early and effective policy interventions.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Information systems** → *Specialized information retrieval*; • **Computing methodologies** → *Feature selection*; Learning linear models; • **Social and professional topics** → Health information exchanges.

KEYWORDS

Disease Surveillance, Big Data, News Analytics, Sparse Signals, Time Series

ACM Reference Format:
Ananth Balashankar, Aashish Dugar, Lakshminarayanan Subramanian, and Samuel Fraiberger. 2019. Reconstructing the MERS Disease Outbreak from News. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '19), July 3–5, 2019, Accra, Ghana.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3314344.3332498

1 INTRODUCTION

The Middle Eastern Respiratory Syndrome - Corona Virus (MERS-CoV) disease is a new illness caused by a type of corona virus found in the Arabian peninsula since 2012. While most corona viruses have only cold-like symptoms, most people with the MERS virus had severe respiratory illness, gastrointestinal problems, sometimes leading to death [9]. As of end of September 2018, there were a total of 2260 laboratory confirmed cases and 803 associated deaths from MERS [20]. Despite the decreasing number of new cases over the years, WHO maintains its global risk assessment as it is mainly acquired from dromedary camels, a popular domesticated animal. There have been 218 instances of exported cases where contact with animals happened in the Middle East, but symptoms later manifest in the home countries of travellers. The difficulty in tracking MERS stems from the fact that, the dromedary camels show no symptoms when they are infected by MERS, making it harder to isolate them.

Early detection of MERS outbreaks is critical for health care resource allocation similar to diseases like malaria, dengue [1] and Ebola [6]. On the ground interventions can be mobilized in a more precise manner if the health agencies understand the local geographic, cultural and socio-economic conditions in a much fine-grained manner. However, structured signals on these aspects are available yearly or quarterly through extensive surveys conducted by organizations like WHO and UNICEF [24], making it difficult to apply traditional machine learning techniques to predict outbreaks, which usually span a few weeks. For communicable diseases specifically, the mobility patterns of people and animals play an important role in determining the risk of an outbreak in a region and measuring this in regions with low access to tracking technology can be non-trivial.

In our study, we measure the factors that impact the propagation of the disease based on their mentions in the news. Specifically, we hypothesize that mobility patterns and access to local health care is impacted due to the presence of conflict within a region. This, in turn, influences the risk of a disease outbreak in a region. We use real time news streams such as GDELT [17] and the Uppsala Conflict data program [13] that aggregate statistics of conflict related death counts within a given geography. We use this localized knowledge in addition to a traditional disease transmission model for MERS [4] which estimates the susceptible, infected and recovered (SIR) number of people in a population based on the instrinsic characteristics of the disease as studied in a hospital. We extract interpretable variables, by running Granger Causality [15] tests

COMPASS '19, July 3–5, 2019, Accra, Ghana

ACM ISBN 978-1-4503-6714-1/19/07…$15.00
https://doi.org/10.1145/3314344.3332498

for each of the hypothesized 56 news based indicators and keep only the ones which are statistically significant. We then embed the trained SIR model with the Granger-causal variables in a multivariate auto-regressive linear model to predict future infected number of cases and deaths.

Using sparse but rich conflict signals from the GDELT news database, our disease outbreak model is able to reconstruct the time series of actual infected cases as reported by WHO with a sum of squared errors which is 3.36x lower than using the standard MERS epidemiological model alone. The news based indicators which are most influential in our model represent the number of people killed, wounded and affected due to conflict in the regions of Lebanon, Kuwait, Egypt and Jordan. Some of these factors negatively influence the population mobility patterns and have disparate influence across regions. In addition to the variations of coefficients for news based factors, we use sensitivity analysis and Granger Causal [15] time lags to interpret how each of these factors affect the timing and scale of the MERS outbreak in the middle east from 2013-2018.

2 RELATED WORK

The environmental, animal and human transmission model [4] provided an understanding of how we could initialize the parameters for the transmission rate in the SIR Model. This work analyzed transmission patterns in a hospital in Saudi Arabia and identified the parameters of the SIR model. Apart from human-human transmissions, this model also incorporates animal-human interaction, especially from dromedary camels which serve as a large reservoir for the transmission of this disease. Incorporating this transmission alongside the human transmission rate significantly improves the accuracy of the model. The WHO currently educates people in the region to stop using animal products which could have come in contact with these camels when an outbreak is imminent.

The Dynamical Transmission Model [26] provided a corroboration to our parameter estimates. The sensitivity analysis provides an overview as to how the parameters would fluctuate on each iteration, which is in line with the modeling based on [4]. These analyses determined the changes that a parameter has on a model and the key drivers in a model(this happens to be the transmission rate b)

News based indicators have been used to predict man-made disasters and other natural events which are worthy of global attention previously in [18]. The tool developed was used to aid journalists in tracking events of consequence from Twitter streams [14]. In our work, we rely on established news sources and their aggregations. Parsing social media feeds would require sophisticated tools to filter false positives and would remain the focus of our future research direction.

Other auxiliary data like internet search history [12] and the web [5] have been used for disease surveillance, but the limitations of a fully unsupervised system without validation can cause spurious correlations as noted in [16]. A more purposeful and dedicated system built for disease tracking have also been deployed in real world systems as shown in [1, 8] rely on time series of structured data collected by specialists who were trained for this specific purpose. In this work, we try to take a combined approach [10, 27] by relying on aggregated news data which is not only easy to scale, but also validated by tools known to journalists and conflict trackers like the Uppsala conflict program. Thus, we aim to extract valid signals from a large news stream corpora to better understand disease transmission properties for MERS.

3 BACKGROUND

In this section, we elaborate on the specifics of the MERS disease and motivate the need of news based modeling to overcome the challenges of addressing sparsity constraints in diseases like MERS. The hypothesis we will motivate in this section is that sparsity of on-the-ground signals relevant to disease modeling can be overcome by augmenting events from news which impact the migration and hence the disease propagation patterns indirectly. Specifically, we explore the scenario where conflict events impact the disease modeling of MERS in the Middle Eastern countries like Saudi Arabia, Kuwait, Lebanon, Egypt is presented here.

3.1 MERS

The Middle East Respiratory Syndrome is a respiratory illness caused by a coronavirus (MERS-CoV) and shows symptoms like fever, cough and shortness of breath. Close to 3-4 people who were infected have died of MERS related complications [9]. Although the disease was first reported in September 2012 in Saudi Arabia, it has since spread across the globe. In 2015, the largest outbreak outside the Arabian peninsula happened in South Korea and was traced back to a traveller from the middle east. MERS symptoms have been varied based on the risk factors like diabetes, heart disease or weakened immune system. While severe complications including pneumonia or kidney failure have led to death, people who have shown milder symptoms or no symptoms have recovered. The incubation period of MERS is usually 5-6 days, but larger variations of 2-14 days have also been observed. This means that people who have come in contact with the virus can show no symptoms for up to 1-2 weeks [9]. This makes detecting MERS extremely difficult as it is known to have been transmitted through close contact with an infected person in addition to infected animals like dromedary camels, a popular animal for transportation in the middle east. Thus, 10 countries in the Arabian peninsula and 17 countries outside it have seen more than 2200 cases of MERS and there continues to be a threat of an outbreak.

3.2 Data Sparsity

As MERS is extremely hard to detect during the incubation period, many patients who show milder symptoms might go untested and can potentially infect people who have a higher risk of developing severe complications. Thus, the number of actual cases of MERS is harder to estimate due to lack of resources for testing and a lack of awareness. Thus, WHO and other health agencies rely on laboratory confirmed cases which form an extremely sparse data source. This will form the ground truth data in our analysis. Since the reports by the disease outbreak team by WHO are carefully cross-checked, it can be weeks or even months before the actual data is available for analysis. The reports are published weekly and sometimes fortnightly on the WHO's website [19] and is released widely. This limits the granularity of our analysis and rules out any real time analysis, daily or less based on streaming signals.

News signals about conflict are also considerably sparse with most coverage in the news relying on local sources that makes aggregation of data time-consuming. Press releases appear in batches, often with aggregate numbers over a longer time window. However, even such sparse news reports capture rich signals of conflict which can be specifically useful to predict the impact on human migration. For example, initial death counts from a conflict gets reported on the day of the event, but the actual numbers are usually updated once more information is learnt and the corresponding statistics are updated. Relying on such sparse corrected sources is much useful than trying to parse all the available data, which can contain false information. We use such rich time series which are curated and verified by journalists and agencies on the ground for our analysis.

Given these sparsity constraints, we aim to reconstruct the time series of the actual number of infected MERS cases based on richer signals extracted from domain-specific knowledge about conflict and the corresponding limited data in the news.

3.3 Disease Outbreak Modeling

Traditional disease outbreak modeling relies on developing a mathematical model which denotes the rates of susceptibility (S), infection (I) and recovery (R) of a disease. This is usually modeled as differential equations where the assumptions are embedded in the way the equations are parameterized [4]. For example, for incurable diseases, recovery (R) is not modeled at all and sometimes, more than one type of infected and susceptible populations are tracked separately based on the mode of disease propagation. These assumptions stem from biological laboratory research which study the intrinsic propagation properties of a disease. Once such a mathematical epidemiological model is constructed by enumerating the number of compartments (S, I, R) and their interactions, its parameters are estimated and validated by a case study of a few specific hospitals and their surrounding regions. A critical parameter estimated through such case studies is the disease's basic reproduction number (R_o), which signifies the risk of the disease becoming an outbreak in a population [25]. An $R_o > 1$, indicates that unless sufficient interventions are not carried out, there would be an exponential increase in the infected population through a multiplicative effect. Mathematically,

$$R_o = \rho(FV^{-1})$$

where ρ is the spectral radius of the next generation matrix, where F is a column matrix denoting the rate of increase in population of compartments and V is the rate of decrease in population from compartments due to all other causes.

3.4 Sensitivity Analysis

Once the parameters of the differential equations are estimated, a thorough sensitivity analysis of the parameters is done to understand how changing any of these parameters affects the susceptible, infected and recovered populations. Mathematically, this is done using the sensitivity index relative to the reproduction number for parameter p [23],

$$SI_{R_o,p} = \frac{\partial R_o}{\partial p} * \frac{p}{R_o}$$

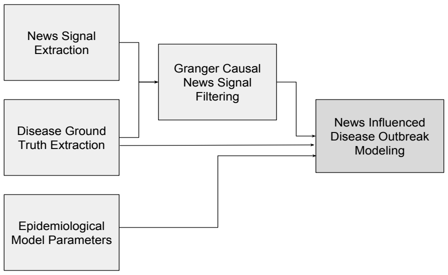
News Signal Extraction
Disease Ground Truth Extraction
Epidemiological Model Parameters
Granger Causal News Signal Filtering
News Influenced Disease Outbreak Modeling

Figure 1: Outline of News-influenced Disease Outbreak Modeling

Higher the $SI_{R_o,p}$, higher is the impact obtained by interventions that influence that parameter. One of the critical assumptions made while such models are used in practice is that the surrounding socio-economic, political and infrastructural environments of the place where the study was conducted and where it is deployed are identical for all matters concerning the spread of a disease. This inherently ignores the changes in the availability of health care and other such extrinsic factors. This deviation of the on-ground reality and conditions of case studies significantly impacts the efficacy of such models. Large data sets of these extrinsic signals are also not readily available in the regions which are most at risk of disease outbreaks.

4 METHODOLOGY

In order to overcome this compounded problem of not being able to scale the epidemiology model to regions which are most at risk, due to the lack of extrinsic knowledge of socio-economic conditions in those fragile states, we resort to the news to extract meaningful signals for disease modeling. However, not all event-indicators in the news are relevant to disease modeling and careful inspection of the variables chosen is required. Hence, we take a conservative approach and filter only those variables related to the factors studied by social researchers for disease outbreak modeling and prescribed by WHO [21]. As per WHO, conflict is the primary factor that increases the risk of spread of infectious diseases like MERS. Hence, early indicators of even such sparse conflict related signals from news streams can significantly boost the accuracy of the SIR model applied for infectious diseases. In the remaining sections, we describe the methodology of our news based models and results.

Building news based models for disease outbreak modeling requires information retrieval tools to extract signals from the news, ground truth data from trusted sources, domain knowledge of the disease captured in graphical models of disease propagation and finally the prediction model which integrates all of this to produce the final estimate of the number of people infected by the disease. This is illustrated in Figure 1.

4.1 News Extraction

In order to incorporate news based signals in our disease outbreak prediction modeling to model extrinsic factors, we need to convert the words present in the news to a suitable representation which can capture the trends in the news. We hence chose to model it as a time series of events and its relevant statistics which are relevant to the disease. For example, for a regional *conflict* which is causing stress, we take into account the number of times that conflict was mentioned in the news and its associated number of deaths, wounded and sickened people. The definition of conflict can be ambiguous depending on the stakeholders and this extraction is conditioned on the domain expertise of the journalists in ensuring that aggregate statistics are not duplicated. These are usually extracted from the news article where it was mentioned. Quite often, the statistics reported are cumulative instead of the incremental change required at time t and hence we needed to build suitable tools and language filters to prune them.

In addition to raw news articles, we also used structured tables which are curated by organizations like Uppsala Conflict Program [22] to extract some of these relevant news signals. These are again suitably filtered using data processing tools. Once these time series were generated, they were normalized such that the time series is centered. This is required so that the variations in raw values across regions are comparable and are not dominated by the largest value. Any time series prediction task does not usually converge unless the time series is stationary and lack seasonal trends. To remove such trends, it is common to take differences until the final time series is stationary. However, in the case of sparse time series where conflict occurs based on seasonal and other trends which are not stationary, we resort to time series chunking. Each time series chunk, denoted by a start and an end date corresponds to a conflict episode and the time series within each episode is ensured to be stationary. We use such time series chunks throughout our prediction task.

4.2 Disease Ground Truth Extraction

Extracting ground truth of the number of cases and deaths associated to a disease can be quite controversial due to differing reports in the news and medical agencies. We rely on trusted sources like the UN, WHO to provide us with these estimates based on on-the-ground healthcare personnel. Some of these trusted sources provide data in the form of monthly reports or bulletins in the form of natural language text. We parse this text and extract relevant statistics like time, number of new cases and deaths reported for a disease across regions. Extraction is done using regular expressions as most of this text usually follows a template, which can be easily reverse-engineered. This provides the time series of the ground truth for the prediction task.

In order to take into account data outages and changes in template, we utilized RSS feeds on the disease outbreak portals to cross check the numbers extracted. These usually serve as efficient notifications of updates, but need to be monitored for changes undetected by web scrapers. Scaling these scrapers to multiple sources and in multiple languages remains out of scope for this task. However, while inspecting the news articles cited in these trusted disease outbreak sources, we usually noted that they were in the local language. Incorporating signals from these would be immensely useful for early-detection of outbreaks.

4.3 Epidemiological Modeling

Disease modeling based on rates of changes in population sizes at different stages of a disease is a common mathematical modeling approach. In this model, populations are compartmentalized and the rate of transfer of individuals from one compartment to another is modeled using differential equations. This can be easily visualized in a graphical model with each node denoting a compartment and the weights of the directed edges denoting the transfer rates. Each compartment is semantically annotated with a stage in their exposure to the disease like "susceptible", i.e sub-population which is at risk of getting the infection, "infected" who have the infection and "recovered" who have either recovered or died from the infection. In all such compartmentalized models, the population of the region is assumed to be constant and transitions between compartments have Markov assumptions.

This makes it easier to denote the graphical model in terms of differential equations with the rates of transfer and the nodes in the model being specific to a disease. MERS being an infectious disease has been studied by epidemiologists and several models have been proposed including SISI and SIR models [26]. SISI model, for example has two types of infections (primary and secondary) in two regions, where primary infections occur from contact with animals and secondary infections occur from contact with other infected humans in hospitals. The corresponding susceptible (S) and infected (I) populations are estimated using links from $S \longrightarrow I \longrightarrow S \longrightarrow I$.

In our modeling, we refer to the SIR (Susceptible, Infected, Recovered) model, a standard mathematical model which predicts how a disease propagates in a closed population over time. It represents the SIR population numbers as a function of time, and describes the time line of an epidemic, by fitting data from case studies on a small number of hospitals in the region where the disease is endemic. The sensitivity of this model is defined by the reproductive number (R_o) and the effect of MERS specific parameters on it are validated by epidemiologists on the population of Saudi Arabia [26]. We can relate the population numbers $s(t)$, $i(t)$ and $r(t)$ by the following differential equations. Solving for $i(t)$, given the initial population numbers, gives us the estimate of number of infected patients, which we refer to as $SIR[t]$ in the following sections.

$$\frac{\partial s}{\partial t} = -bs(t)i(t) \tag{1}$$

$$\frac{\partial i}{\partial t} = bs(t)i(t) - ki(t) \tag{2}$$

$$\frac{\partial r}{\partial t} = ki(t) \tag{3}$$

where b = rate of transmission, k = rate of recovery

4.4 Granger Causal Testing

Given two time-series X and Y , the Granger causality test checks whether the X is more effective in predicting Y than using just Y and if this holds then the test concludes X "Granger-causes" Y [15]. However, if both X and Y are driven by a common third process with

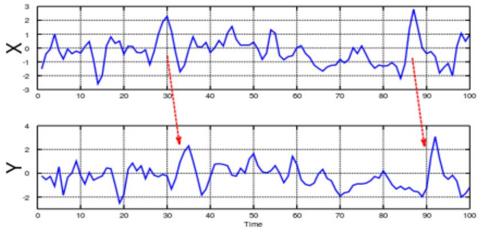

Figure 2: Granger causal link between two time series

different lags, one might still fail to reject the alternative hypothesis of Granger causality that X "does not Granger-cause" Y. Hence, in our modeling, we explore the possibility of causal links ignoring confounding variables due to the domain knowledge that there are no such confounding variable noted by the WHO. We note that if such an unobserved confounding variable exists, it is not considered in our Granger causality test.

In order to ensure that the news variables chosen are indeed related to the disease outbreak and not spurious correlations, we ran the Granger Causal test [15] between all of the news indicator variables (x) and the disease outbreak (y) as seen in Figures 2. We chose linear equations as our choice of modeling the prediction between x and y as it retains the benefits of interpretability in their coefficients. If, m, p, q denote the time lags of y, x in the auto-regressive equation at time t, then we can write:

$$y_t = a_0 + a_1 y_{t-1} + \ldots + a_m y_{t-m} + b_p x_{t-p} + \ldots + b_q x_{t-q} + error_t$$

Specifically, x is known to Granger-cause y, if there exists at least one non-zero coefficient of x which then leads to a significant improvement in prediction error over the case when we just use lagged values of y. We perform parametric F-tests on the non-zero coefficients of lagged variables and chose only the significant variables (p-values ≤ 0.05) to reject the null hypothesis that "the news indicator variable (x) does not Granger-cause the disease outbreak (y)". The chosen Granger Causal news variables are denoted by the vector $News[t]$ for a given time t, in the next sections. Note that since true causality is hard to establish through observational studies, our goal here is to only find news variables which depict "predictive causality" and better predict future time series of the disease outbreak.

4.5 News Influenced SIR Modeling

Incorporating the news signals which are Granger Causal of the disease outbreak infections, into the epidemiological SIR model is the main methodological contribution of the paper. One option is to make changes to the equilibrium of the SIR model by altering the nodes in the graphical model and estimating the corresponding changes based on compartments induced by the news variables. This however does not scale to every disease specific model. Reconfiguring the disease model directly requires a lot of domain knowledge of both the disease and the related news variable, and remains out of scope of our paper.

Instead, we perceive the SIR model as yet another time series variable in a multivariate linear regression. This makes it possible to model other diseases easily in a similar manner without having to worry about the complex differential equations that govern the epidemiological transmission model of each disease. Now that we have the relevant news conflict variables chosen by the Granger Causal Test $News[t]$ and the MERS SIR model's value $SIR[t]$, we train a multivariate auto-regressive model with Lasso penalty [2] using glmnet [11] from lagged values of the ground truth $I_{t-\delta}$ and the regression variables as follows, where A, B, C are weight matrices, maximum lag δ, and for any matrix x, let $x_{t-\delta} = x[t-\delta : t-1]$.

$$I[t] = A.I_{t-\delta} + B.News_{t-\delta} + C.SIR_{t-\delta} \tag{4}$$

$$\underset{A,B,C}{min} \; \|I[t] - A.I_{t-\delta} - B.News_{t-\delta} - C.SIR_{t-\delta}\|_2^2 \tag{5}$$

$$\text{subject to} \|(A, B, C)\|_1 \leq r, \text{for a Lasso penalty } r. \tag{6}$$

The non-zero news coefficients that remain in the Lasso equation best explain the difference between SIR and ground truth in the News influenced disease model (Figure 5). The Lasso regularization embodies a variable selection procedure that ensures that only the most important variables are selected for prediction. We also reduce collinear variables in order to ensure that the Lasso regularizer does not pick variables which depict the same underlying event. This can be seen as a pre-processing step of removing a potential confounder variable as we cannot remove it once the regression model is trained. We use Variance Inflation Scores to prune out collinear variables [3]. This ensures that only those variables which cannot be estimated using the remaining news variables are used in the prediction task.

5 EVALUATION

In this section, we explain the datasets used and the implementation details in the news influenced disease models.

5.1 Dataset

In this section, we describe the disease outbreak ground truth source and the news event databases used to extract conflict related signals in the region.

5.1.1 WHO-UN Dataset. The WHO-UN website [20] presents a collection of articles, which are updated every 8 to 15 days. Articles on each disease include statistics such as the number of cases or deaths and the date of the detected disease. The total size of this data spans 400 events for 192 countries from 2013 to 2018. There are 242 articles mentioning MERS, with breakdown of aggregate cases for each of the 12 Middle eastern regions + South Korea (traced back to a traveler from the Middle East). This data serves as our ground truth set.

5.1.2 GDELT Dataset. GDELT 1.0 Global Knowledge Graph [17] monitors the world's news from every country in over 100 languages with more than 1.5 billion events per year from April 2013 to Jan 2018, updated daily. These events are categorized based on killings or other crises such as natural disasters. It also provides a daily human count for each of these event types from sources like AFP, BBC monitoring, AP, WP, NYT and aggregator tools like Google News. We particularly focus on *killed, wounded, sickened and affected* events reported in each of the 12 regions as shown in Figure 3.

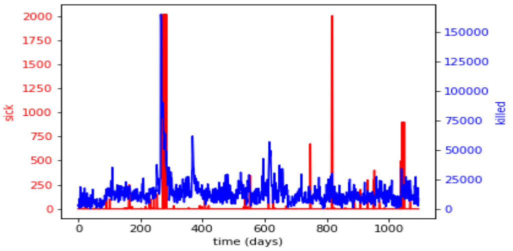

Figure 3: Killed and Sickened count in GDELT

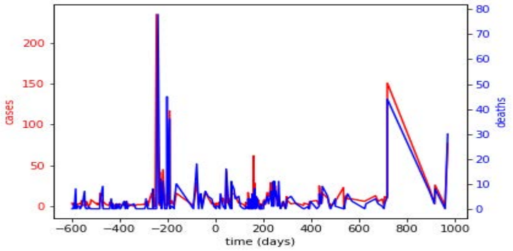

Figure 4: Number of new cases and deaths by MERS as reported by UN-WHO (> 70% in Saudi Arabia)

5.1.3 Uppsala Conflict Data Program. The Uppsala Conflict dataset [13, 22] provides deaths from organized violence keyed by a conflict ID and country, where each conflict has at least 25 related deaths in a year. The data set is presented as a time series with an yearly number of deaths per conflict. We focused on 8 of the 12 MERS regions which had a conflict (includes Saudi Arabia).

5.2 Data Preprocessing

We retrieved all the disease outbreak news articles from the UN website. These were later filtered to contain only the headline, timestamp, new cases and deaths using rule based string matching as can be seen in Figure 4. We extracted time series for each of the 48 normalized news indicator variables to range in $[-1, 1]$ for all (country, event-type) tuples from GDELT and 8 conflict variables per country from Uppsala. Time series chunking is also done to ensure that all the time series used for a specific time window is stationary. We take differences between consecutive values until stationarity is achieved. If we do not observe stationarity after differencing twice, we drop that time series from consideration as it no longer holds any interpretable meaning.

5.3 Model Parameters

The values of the SIR model's parameters as noted in Eqns [1-3] are predetermined. Specifically, the transmission rate b = *1.4248* and recovery rate k = *0.1484*, used are based on the epidemiology study for MERS done in [4], as opposed to making theoretical estimations. The maximum time lag used for Granger Causal link estimation $\delta = 6$ weeks. The same maximum time lag was also used for the final news influenced multivariate auto-regressive model. This value was chosen based on the minimum size of the time chunk obtained in the data for 20 weeks.

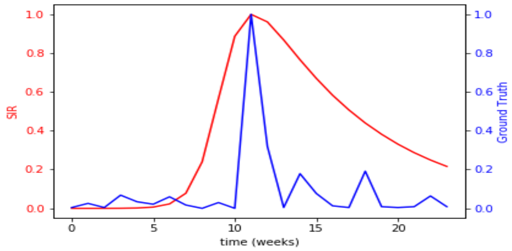

Figure 5: Difference between SIR and ground truth for an outbreak time window.

5.4 Implementation

An overview of our implementation of building a news influenced disease model is given in Algorithm 1.

Algorithm 1: News Influenced Disease Modeling

```
1 Extract the ground truth timeseries for number of cases from
    the WHO-UN articles
2 Fit the epidemiological SIR model using pre-defined MERS
    specific parameters
3 Filter relevant conflict signals from GDELT and Uppsala by
    running Granger Causality tests
4 Train a multivariate auto-regressive model with SIR estimate
    and relevant conflict signals
```

6 RESULTS

In this section, we will discuss the performance of the News Influenced disease model against several baselines. We pick 10 short outbreaks from 2013-2018, each spanning 21 weeks with the peak of the outbreak in the middle of the time series. The disease numbers reported are new cases and new deaths reported per week due of the disease. We fit the SIR model for each of these 10 outbreaks as per the variables mentioned above. We then normalize both the ground truth values and the SIR modeled values such that minimum and maximum values in the time series are scaled between 0 and 1 as seen in Figure 5. The final error calculated is the sum of the point-wise (one point per week) squared errors between the modeled and the ground truth. We report the average 10-fold cross validation error across multiple outbreaks.

6.1 Choice of News Source

In building a news based disease model, the source of the signals incorporated can have a significant impact on the trustability and accuracy of a model. Choosing between news sources can also influence the implementation requirements if this model were to be scaled. We tried various sources for the $News_{t-\delta}$ variable in Equation 4: 1) Conflict signals from GDELT 2) Conflict signals from Uppsala 3) Both GDELT and Uppsala 4) Only GDELT signals (no

SIR, Uppsala). As mentioned in Table 1, SIR model with GDELT signals performs the best, reducing the error from the baseline SIR model of 8.99 to 2.68, an improvement by a factor of 3.36. The results presented in Table 1 are average errors from 10-fold cross validation of the episodes identified from time chunking. The low standard deviation of the errors shows that there is not a huge variation based on which chunks of outbreak episodes were used for training, indicating consistency and internal validity of the news influenced disease model.

Uppsala conflict signals were not useful for predicting the disease outbreak time series. We attribute this to the hand curated condensed extremely sparse (yearly) representation in the Uppsala event database. GDELT on the other hand is a daily aggregated database which captures the signals as represented in the news. This shows that GDELT has a better trade-off between aggregation coarseness and the time duration taken to put out verified conflict statistics. Another surprising result was that, using factors from GDELT alone in the multivariate auto-regressive prediction, produces a much lower error than the SIR model. This clearly indicates that local environmental and social factors are as important if not more important than the propagation properties of the disease within hospitals.

Model	Residual sum of squares error	Std deviation
SIR	8.99	0.65
SIR+GDELT	**2.68**	**0.42**
SIR+Uppsala	35.30	4.34
SIR+GDELT+Uppsala	2.81	0.43
GDELT	5.08	1.29

Table 1: Performance of News Influenced SIR Model

In Table 2, we see that , the news influenced SIR model performs well across outbreak episodes. The results presented are for those cross-validation rounds when the said episode was used for testing. The low variation is indicative that we can use the approach in predicting future outbreaks and consistently explain the factors that were highlighted in the model.

Outbreak Episode	RMSE
March–June 2014	2.06
July–December 2014	0.40
January–April 2015	1.29
May–June 2015	2.38
June–July 2015	6.06
July–Sep 2015	1.92
April 2016 – August 2017	1.63

Table 2: Performance of News Influenced SIR Model across Episodes

6.2 Explainability of News Signals

Claiming lower prediction errors for the disease transmission patterns is not useful unless the model can be explained in terms of the multiple conflict signals in our model. Since, the time series used for analyzing each episode are normalized, we can directly compare the values of the coefficients. We chose the coefficients with the maximum absolute value over the many cross validation runs. This is highly correlated to the sensitivity index (SI_{R_o}) usually computed for disease outbreak models. The sign of the coefficients also indicate how conflict might indirectly influence the transmission patterns of the disease outbreak as can be seen in Table 3. In addition to the raw value of the coefficient, it is also useful to determine what is the expected time lag between a news signal appearing in the news and the expected influence on the number of infected people. This number (in weeks) when combined with the coefficient value, provides the estimate of when and how much of an impact a signal in the news will have on the disease outbreak.

To illustrate this explainability, we choose to analyse the model predicting the outbreak from March-June 2014. Table 3 shows that although the time-lagged ground truth (actual counts from WHO) and SIR model remain the most important variables, conflict signals like kills in Kuwait and Lebanon (neighboring regions to Saudi Arabia) have a negative impact on the transmission of the disease, whereas increase in wounded and sick people in Egypt and affected people in Jordan indicate the increase in disease transmission of MERS. While events related to people being killed in conflicts could be traced to severe restriction of migration, while events related to being affected or wounded could seen as early indicators of people migrating due to the upcoming severe conflicts. While we note that there might be some feedback built into our model based on sick events retrospectively used, this requires further explorations.

Feature	Coefficient	Best time lag (weeks)
Lagged_Truth	0.17	1
SIR	0.23	3
kill_Kuwait	-0.17	5
kill_Lebanon	-0.15	5
wound_Egypt	0.12	5
affect_Jordan	0.10	1
sick_Egypt	0.03	1

Table 3: Important factors of the News Influenced SIR model

6.3 Implications

The above results which show more than 3x reduction in root means squared error is significant also because of the evidence it provides confirming the hypothesis articulated by WHO that conflict causes severe distress and exacerbates the spread of diseases. All the coefficients reported above are statistically significant (p-values < 0.05). Additionally, the time lags corresponding to each of the variables in multivariate regression provides us actionable information to facilitate timely interventions for disease containment. For example, when people in Jordan were affected by conflict in March–June 2014, it led to an increase in MERS infected cases due to migration 1 week after the said conflict as illustrated in Table 3. Similar insights can be extracted for other outbreak episodes too. As per the current WHO fact sheet about MERS [20], there is no vaccine available for MERS, but appropriate hygiene needs to be practiced by people handling dromedary camels and the consumption of raw animal products should be minimized during the outbreak. Such advice is particularly useful for people affected in the region as symptoms of MERS appear later in the infection stage and is not easily distinguishable by health care workers. This early warning

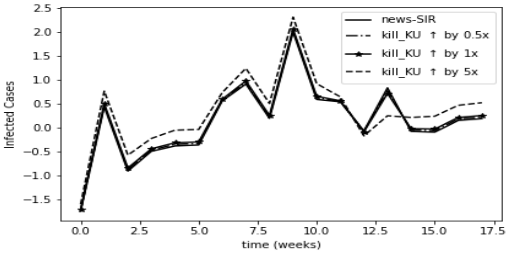

Figure 6: Sensitivity analysis based on number killed in conflicts in Kuwait indicates a uniform value shift in number of infected cases in March-June 2014.

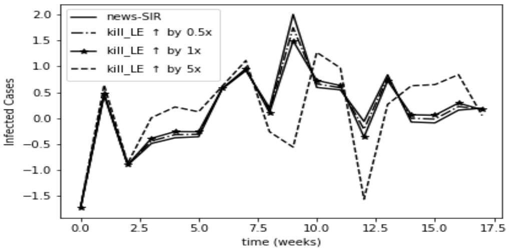

Figure 7: Sensitivity analysis based on number killed in conflicts in Lebanon indicates a disparate phase and value shift in number of infected cases in March-June 2014.

indicator is also beneficial for health care workers to prepare and use appropriate eye protection and other containment strategies including proactive blood tests.

6.4 News Sensitivity

Along with the timeliness of the news based disease model, we can also measure the sensitivity of the model for changes in the future related to conflict. This provides a way to distinguish the variations in the disease propagation pattern with any future significant escalation in conflict. We illustrate this sensitive analysis on the MERS outbreak from March–June 2014. Similar analyses can be done on other outbreak episodes too. We observe that even though some coefficients of news based variables are closer in value (kill_Kuwait and kill_Lebanon), the patterns they depict with respect to sensitivity significantly vary due to the underlying time series. For example, in Figure 6, we mostly see an increase in the number of MERS infected cases throughout the time series uniformly with increase in the number of people killed due to conflict in Kuwait. Whereas,in Figure 7 for Lebanon, we see both a phase shift and change in number of MERS infected cases with increase in number of people killed in conflict. We correspondingly see specific time periods where the impact is the highest from the conflict (week 9) as can be seen in Egypt for number of people wounded in conflict in Figure 8. Such variations in expected number of infected cases was not previously known or understood through time-based sensitivity analysis. This not only allows decision makers to categorize different types of conflicts, but also increases the awareness of the complexity and tight linkage between conflict and disease outbreaks.

7 DISCUSSION

Is news a good modeling choice?: There is usually a disconnect between the disease modeling and the health care policy communities. While, the former relies on mathematical modeling to extract the most accurate parameters of the model, the latter cares more about adapting to on-the-ground realities and incorporating information on-the-fly into decision making. Mathematical models which are rigid and harder to interpret is usually not implemented by policy decision makers. This has led to customized web-tools built for practitioners to interface their knowledge with the underlying model [7]. We are inspired by such approaches and extend it to directly incorporate local information from the news. While news based modeling has the potential pitfall of relying on sentiments more than facts, we incorporate verified statistics about conflicts which get reported instead of the story around the event, which can be interpreted subjectively. This makes news based modeling a worthwhile choice for disease outbreaks which communicable through social contact.

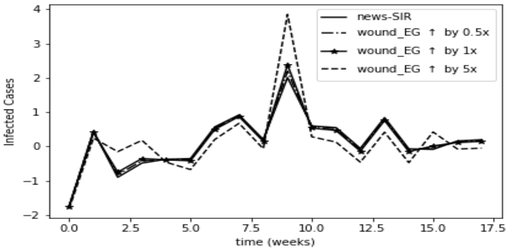

Figure 8: Sensitivity analysis based on number wounded in conflicts in Egypt shows varied shifts in number of infected cases at specific time intervals in March-June 2014.

Is MERS different than other diseases?: As MERS is heavily localized to countries in the Arabian Peninsula, it makes local news based modeling easier and drastically reduces the scope of news articles to be studied. MERS also has the clear distinction of a disease which spreads due to human and animal transportation in this region. This movement of people, animals and products is known to be a social indicator of the underlying political, economic and humanitarian conditions in the region. Thus, modeling MERS through news based modeling in the middle east makes more sense than other vector borne diseases or in any other region, outside the area of impact of the above macro-level events like conflict.

How can this be used at scale?: Having been able to reconstruct the time series of previous episodes of MERS with low average prediction error and low deviation in errors across all cross-validation of episodes, it provides us confidence to incorporate this model to predict future outbreaks. The model however might have to be tweaked to account for the efficient implementation of health care advisories issued by the WHO, which has significantly reduced the risk of MERS since it first occurred in 2012. This would impact the SIR component of the news influenced model, but not the factors learnt from the news, which are updated by design. Such a model, if adopted by the WHO or other health agency can significantly improve prediction of disease outbreaks based on historical patterns in the news, and lead to better intervention and information dissemination strategies.

When would this model not work?: Further analysis however is required to breakdown the different types of conflict and the corresponding regions they impact. This can be done through spatial and text based categorization of the news articles which mention conflict. Such a model would however significantly suffer from the sparsity in the data post the categorization of conflict, similar to how news signals from the Uppsala Conflict Data program proved to be less effective due to the sparsity of data. This challenge needs further model improvements and cannot be addressed by the current news based model. One option we are actively pursuing is the tree-based factorization of the news signals which combines the best of both sub-categorization and larger datasets in a hierarchical approach.

8 CONCLUSION

Susceptibility, infection and recovery is modeled in disease transmission models using intrinsic properties of the disease. However, extrinsic factors also influence disease transmission and have been previously unexplored. We study the effect of regional conflict on the mobility patterns of people and animals for the transmission of MERS-CoV and show that by augmenting conflict based signals in real time news streams with a standard MERS SIR model, we significantly lower the infected population prediction error. Inspection of our news influenced disease model provides a human interpretable understanding even with very sparse signals.

REFERENCES

[1] Nabeel Abdur Rehman, Shankar Kalyanaraman, Talal Ahmad, Fahad Pervaiz, Umar Saif, and Lakshminarayanan Subramanian. 2016. Fine-grained dengue forecasting using telephone triage services. *Science Advances* 2, 7 (2016). https://doi.org/10.1126/sciadv.1501215 arXiv:http://advances.sciencemag.org/content/2/7/e1501215.full.pdf

[2] Andrew Arnold, Yan Liu, and Naoki Abe. 2007. Temporal Causal Modeling with Graphical Granger Methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*. ACM, New York, NY, USA, 66–75. https://doi.org/10.1145/1281192.1281203

[3] D. A. Belsley, E. Kuh, and R. E Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.*

[4] Gerardo Chowell, Seth Blumberg, Lone Simonsen, Mark A. Miller, and CÃľcile Viboud. 2014. Synthesizing data and models for the spread of MERS-CoV, 2013: Key role of index cases and hospital transmission. *Epidemics* 9 (2014), 40 – 51. https://doi.org/10.1016/j.epidem.2014.09.011

[5] Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. 2008. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* 24, 24 (2008), 2940–2941. https://doi.org/10.1093/bioinformatics/btn534

[6] Sam Crowe, Darren Hertz, Matt Maenner, Ruwan Ratnayake, Pieter Baker, R Ryan Lash, John Klena, Seung Hee Lee-Kwan, Candice Williams, Gabriel T Jonnie, Yelena Gorina, Alicia Anderson, Gbessay Saffa, Dana Carr, Jude Tuma, Laura Miller, Alhajie Turay, Ermias Belay, and Centers for Disease Control and Prevention (CDC). 2015. A plan for community event-based surveillance to reduce Ebola transmission - Sierra Leone, 2014-2015. *MMWR. Morbidity and mortality weekly report* 64, 3 (January 2015), 70âĂŤ73. http://europepmc.org/articles/PMC4584562

[7] Ashlynn R. Daughton, Nicholas Generous, Reid Priedhorsky, and Alina Deshpande. 2017. An approach to and web-based tool for infectious disease outbreak intervention analysis. *Nature Scientific Reports, volume 7, Article number: 46076* (2017).

[8] Gunther Eysenbach. 2009. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *J Med Internet Res* 11, 1 (27 Mar 2009), e11. https://doi.org/10.2196/jmir.1157

[9] National Center for Immunizations and Division of Viral Diseases Respiratory Diseases. 2018. Middle East Respiratory Syndrome (MERS). (May 2018). https://www.cdc.gov/features/novelcoronavirus/index.html

[10] Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Brownstein. 2008. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association* 15, 2 (2008), 150–157. https://doi.org/10.1197/jamia.M2544

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33, 1 (2010), 1–22. http://www.jstatsoft.org/v33/i01/

[12] Matthew S Patel Rajan Brammer Lynnette Smolinski Mark Brilliant Larry Ginsberg, Jeremy Mohebbi. 2008. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature. 457. 1012-4. 10.1038/nature07634* (2008).

[13] Nils Petter Gledistch, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Havard Strand. 2002. Armed Conflict 1946-2001: A New Dataset. *Journal of Peace Research* 39, 5 (2002), 615–637. https://doi.org/10.1177/0022343302039005007 arXiv:https://doi.org/10.1177/0022343302039005007

[14] Janaína Gomide, Adriano Veloso, Wagner Meira, Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. 2011. Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. In *Proceedings of the 3rd International Web Science Conference (WebSci '11)*. ACM, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/2527031.2527049

[15] Clive Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* 37, 3 (1969), 424–38. https://EconPapers.repec.org/RePEc:ecm:emetrp:v:37:y:1969:i:3:p:424-38

[16] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, 6176 (2014), 1203–1205. https://doi.org/10.1126/science.1248506 arXiv:http://science.sciencemag.org/content/343/6176/1203.full.pdf

[17] Kalev Leetaru and Philip A. Schrodt. 2013. GDELT: Global data on events, location, and tone. *ISA Annual Convention* (2013).

[18] Armineh Nourbakhsh, Quanzhi Li, Xiaomo Liu, and Sameena Shah. 2017. "Breaking" Disasters: Predicting and Characterizing the Global News Value of Natural and Man-made Disasters. *CoRR* abs/1709.02510 (2017). arXiv:1709.02510 http://arxiv.org/abs/1709.02510

[19] World Health Organization. 2018. WHO MERS Disease Outbreak News. (August 2018). http://www.who.int/csr/don/archive/disease/coronavirus_infections/en/

[20] World Health Organization. 2018. WHO MERS Global Summary and Assessment of Risk, Disease Outbreak News. (August 2018). http://www.who.int/csr/disease/coronavirus_infections/risk-assessment-august-2018.pdf

[21] World Health Organization. 2019. Conflict and Infectious Diseases. (March 2019). https://www.who.int/tdr/research/social_research/conflict/en/

[22] TherÃľse Pettersson and Kristine Eck. 2018. Organized violence, 1989-2017. *Journal of Peace Research 55* (2018).

[23] A. Saltelli, K. Chan, and E. M. Scott. 2000. *Sensitivity Analysis. Wiley Series in Probability and Statistics.*

[24] UNICEF. 2015. Statistics and Monitoring: Country Statistics. (August 2015). https://www.unicef.org/statistics/index_countrystats.html

[25] P. van den Driessche and James Watmough. 2002. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences* 180, 1 (2002), 29 – 48. https://doi.org/10.1016/S0025-5564(02)00108-6

[26] Benny Yong and Livia Owen. 2016. Dynamical transmission model of MERS-CoV in two areas. *AIP Conference Proceedings* 1716, 1 (2016), 020010. https://doi.org/10.1063/1.4942993 arXiv:https://aip.scitation.org/doi/pdf/10.1063/1.4942993

[27] Victor L. Yu and Lawrence C. Madoff. 2004. ProMED-mail: An Early Warning System for Emerging Diseases. *Clinical Infectious Diseases* 39, 2 (2004), 227–232. https://doi.org/10.1086/422003 arXiv:/oup/backfile/content$_{p}ublic/journal/cid/39/2/10.1086_{4}22003/3/39-2-227.pdf$