Title: *A Generalizable Data Assembly Algorithm for Infectious Disease Outbreaks*

Authors: Maimuna S. Majumder^{1,2} and Sherri Rose³

Affiliations: [1] Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, 02115 [2] Department of Pediatrics, Harvard Medical School, Boston, MA, 02115 [3] Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA 94305

Corresponding Author Information:

Name: Maimuna S. Majumder

Email Address: maimuna.majumder@childrens.harvard.edu

Mailing Address: 300 Longwood Avenue, Boston, MA 02115

Word Count (Text Only): 1736

Abstract:

Background & Objective: During infectious disease outbreaks, health agencies often share text-based information about cases and deaths. This information is usually text-based and rarely machine-readable, thus creating challenges for outbreak researchers. Here, we introduce a generalizable data assembly algorithm that automatically curates text-based, outbreak-related information and demonstrate its performance across three outbreaks.

Methods: After developing an algorithm with regular expressions, we automatically curated data from health agencies via three information sources: formal reports, email newsletters, and Twitter. A validation data set was also curated manually for each outbreak.

Findings: When compared against the validation data sets, the overall cumulative missingness and misidentification of the algorithmically curated data were $\leq 2\%$ and $\leq 1\%$, respectively, for all three outbreaks.

Conclusions: Within the context of outbreak research, our work successfully addresses the need for generalizable tools that can transform text-based information into machine-readable data across varied information sources and infectious diseases.

Introduction.

Since 2000, thousands of infectious disease outbreaks have been reported by the World Health Organization (WHO) globally [1]. A considerable subset of these have been due to emerging zoonotic pathogens, including the novel coronavirus SARS-CoV-2, the causative agent of the Coronavirus Disease 2019 (COVID-19); its predecessors, Middle East Respiratory Syndrome (MERS) coronavirus and SARS-CoV-1; Zika virus, and Ebola virus, among others [2–4]. Emergence of these pathogens has been driven by the increasing permeability of the animal-human interface, whereas ease of travel has enabled their transmission across borders [5,6]. Not all outbreaks from the last two decades have been due to emerging infections, however; notably, due to increasing vaccine hesitancy around the world, *re-emerging* diseases, such as measles and mumps, have experienced a resurgence as well [7,8].

During these outbreaks, epidemiological information from a variety of data sources—from formal reports by the WHO to email newsletters and social media posts from national ministries of health—is often made available to the public, including researchers responsible for monitoring and mitigation efforts [1,9–14]. Unfortunately, these publicly available data are typically locked in blocks of text that are rarely machine-readable [15], which poses a considerable roadblock for surveillance and response activities that hinge on mathematical modeling (e.g., data-driven allocation of ventilators or vaccines). To overcome this hurdle, researchers typically commit substantial labor towards manually curating and converting these text-based data into an analyzable format (e.g., comma-separated values, CSV). The time and effort required is often directly related to the complexity of the available information.

In this paper, we introduce a generalizable data assembly algorithm to automate curation of text-based, outbreak-related information and demonstrate its performance across three recent case study outbreaks: measles in Samoa (2019), Ebola in the Democratic Republic of the Congo (DRC) (2018–2019), and MERS in South Korea (2015). We implement this algorithm on source text of increasing complexity from social media (i.e., Twitter), email newsletters, and WHO disease outbreak news (DON) reports, respectively, to produce machine-readable CSV files for each of our three case studies. Though the data available for curation vary across source texts, the underlying structure of the algorithm—regular expressions to extract pertinent outbreak-related information—remains constant across applications and is generalizable.

The source texts considered in this study represent a spectrum of information complexity, and when combined with mathematical modeling approaches, can be used to inform decision-making during infectious disease outbreaks. For measles in Samoa and Ebola in the DRC, we extract simple aggregate statistics (e.g., case counts) over time, which can be used for case count projections, assessment of intervention performance, and vaccination rate estimation [17–30]. Meanwhile, for MERS in South Korea, we extract more complex multi-feature patient-level data (i.e., data in which every row is a patient, and every column is a feature), which enable reconstruction of transmission networks and evaluation of risk factors associated with mortality [31–39].

Methods.

Data on the evolving epidemiology of each outbreak were first manually curated for validation purposes. Summary information for each study is available in Table 1. Aggregate

cases and deaths associated with the measles outbreak in Samoa were collected from the Government of Samoa Twitter account from November 22, 2019 (date of first tweet) to December 8, 2019 (date of last tweet) [9,10]. Similar aggregate statistics were also collected for the Ebola outbreak in the DRC from email newsletters issued by the Ministère de la Santé RDC (MSRDC) from August 6, 2018 (date of first newsletter received) to July 31, 2019 (date of last newsletter received) [11,12]. Finally, patient-level data were collected from WHO DON reports for the MERS outbreak in South Korea from May 30, 2015 (date of first report) to June 9, 2015 (date of last report) [13,14]. These same text-based data were then algorithmically collected using our data assembly algorithm.

The assembly algorithm was developed in the Python programming language and, as shown in Figure 1, uses regular expressions and trigger phrases to automatically transform semi-structured text-based information into machine-readable data. Here, trigger phrases are the phrases that accompany the information of interest in a given block of text. When these phrases are translated into searchable patterns of characters (i.e., regular expressions), they act as “triggers” for the data assembly algorithm to identify and collect information for the desired fields (i.e., variables). This underlying regex-based structure enables generalizability of the algorithm to a wide variety of source texts and information types, as demonstrated by the three case study outbreaks selected.

For the measles case study, the following three data fields were automatically curated using our assembly algorithm: cumulative cases, incident cases, and cumulative deaths. Seventeen rows of data, where each row is a date, were collected across these three fields for a total of 51 cells. Similarly, data for the following 10 fields were automatically curated for the Ebola case study: confirmed cumulative cases, total

cumulative cases (confirmed + probable), confirmed cumulative deaths, total cumulative deaths (confirmed + probable), cumulative cases recovered, cumulative vaccinations deployed, cumulative vaccinations deployed in Region A, cumulative vaccinations deployed in Region B, cumulative vaccinations deployed in Region C, and cumulative vaccinations deployed in Region D. Across these 10 fields, 360 rows of data, where again each row is a date, were collected for a total of 3600 cells. Finally, data for the MERS case study were automatically curated to populate the following five fields: documented sex, age, date of symptoms, date of diagnosis, and healthcare worker status. Sixty-three rows of data, where each row is a patient, were collected for a total of 315 cells across these five fields.

In all three case study outbreaks, the manually curated data for the aforementioned fields were used to validate the performance (i.e., missingness and misidentification) of the assembly algorithm. Missingness is defined as a cell for which the algorithm did not curate a value but for which a value was available when compared against manual curation. Misidentification is defined as a cell for which the algorithm curated a value but for which the value was incorrect when compared against manual curation. Given its intended application in outbreak settings, the assembly algorithm was designed conservatively, placing priority on increasing accuracy over decreasing missingness. Code for all three implementations of the assembly algorithm, as well as the manually collected validation data, are available at https://github.com/mmajumder/Data_Assembly_Algorithm.

Results.

When validating algorithmically collected data against manually collected data, the data assembly algorithm performed well for all three iterations. Across the entirety of each

outbreak reporting period, overall cumulative missingness for the case studies was 0% (0 cells) for measles, 1% (34 cells) for Ebola, and 2% (7 cells) for MERS, while overall cumulative misidentification was 0% (0 cells), 0% (0 cells), and 1% (3 cells), respectively.

Because the reporting period for the Ebola outbreak was considerably longer (368 days) than the measles (16 days) and MERS (11 days) case studies, we also examined missingness and misidentification over time by day for the Ebola case study. Notably, the assembly algorithm exhibited steady gains in cumulative accuracy from August 2018 through June 2019, as displayed in Figure 2. Decreased cumulative availability of data in the source itself (i.e., fields for which MSRDC reported data in May 2019 but no longer reported in June 2019) coincided with minor decreases in cumulative accuracy between June 2019 and August 2019. Cumulative missingness dropped from 5% in August 2018 to near 0% in August 2019, and due to the conservative nature of the assembly algorithm, cumulative misidentification was 0% over the same time period.

Discussion.

By showcasing its performance within the context of three distinct infectious disease outbreaks, we demonstrated the generalizability of our data assembly algorithm across diverse source texts and information types. Intuitively, we found that algorithmic curation of more complex data (e.g., multi-feature patent-level data for MERS in South Korea) exhibited slightly higher rates of missingness and misidentification than simpler data (e.g., case counts over time); however, overall cumulative performance for both metrics was impressive across curated fields for all three case study outbreaks.

The fields for which data were automatically curated by our assembly algorithm were selected purposefully given their long-standing utility to mathematical modeling for informed epidemiological decision-making. Historically, counts of cases and deaths over time—fields that were collected both for measles in Samoa and for Ebola in the DRC—have been used to model the transmission dynamics associated with outbreaks, including important epidemiological parameters such as fatality rates and reproduction numbers [17–30,40–45]. These parameters are critical to formulating case count projections [17–21] and assessing performance of interventions [22–25], which enable public health decision-makers to approach outbreaks from a position of preparedness. Furthermore, these parameters can also be used to model vaccination rates during outbreaks of vaccine-preventable diseases, which can be leveraged to lobby for the resources necessary to vaccinate vulnerable communities [26–30]. Meanwhile, patient-level “line list” data have traditionally been employed to assess risk factors for different outcomes [31–38]; indeed, the data presented in this paper for MERS in South Korea have been used precisely in this way to assess risk factors for mortality given MERS-CoV infection [31,32], as well as for transmission to others following infection [38]. Such analyses allow for improvements to resource allocation both with respect to patient care (i.e., preferentially allocate intensive care units to patients who are less likely to survive infection) and with respect to contact-tracing (i.e., preferentially allocate resources to contact trace individuals who are more likely to transmit to others following infection), among other applications.

As recently noted by George et al. [15], tools that can transform text-based information into machine-readable data are urgently needed by the outbreak management community. Given the epidemiological utility of the data types curated by our data

assembly algorithm across our three case study outbreaks, we believe that the usefulness of the work we present here will persist as infectious diseases continue to emerge and re-emerge. We encourage other researchers to apply it to novel contexts (i.e., new outbreaks), while carefully considering the ethical implications before deployment in new settings [46]. Our algorithm is designed to generalize across diseases and enable the democratization of essential epidemiological data that are otherwise locked in blocks of non-machine-readable text. However, despite strong accuracy and missingness assessments for all three case study outbreaks considered in this paper, we recommend that random manual checks be implemented to validate the robustness of our data assembly algorithm when employed during future outbreaks.

178 **Tables.**

179

180 Table 1. Data collected across case study outbreaks.

Case Study	Data Source	Reporting Period	Number of Fields	Total Cells Curated
Measles in Samoa	Twitter	November 22, 2019 – December 8, 2019	3	51
Ebola in the DRC	Email Newsletters	August 6, 2018 – July 31, 2019	10	3600
MERS in South Korea	Disease Outbreak News Reports	May 30, 2015 – June 9, 2015	5	315

181

Figures.

Date	Raw Text	Assembly Algorithm Excerpt	Output
6-Aug-18	Au total, 43 cas de fièvre hémorragique ont été signalés dans la région, dont 16 confirmés et 27 probables.	<pre>[...] def get_field_three(mytext): regex = r"(?:le cumul des cas est de) ([+-]?[0-9]*[.]?[0-9]+) ([+-]?[0-9]*[.]?[0-9]+) (?:cas de fièvre hémorragique)" return(next(string for string in pageiterator(regex,mytext)[0] if string).replace(".",",")) [...]</pre>	43
...	...		...
11-Jun-19	Depuis le début de l'épidémie, le cumul des cas est de 2.071, dont 1.977 confirmés et 94 probables.		2071
12-Jun-19	Depuis le début de l'épidémie, le cumul des cas est de 2.084, dont 1.990 confirmés et 94 probables.		2084

Figure 1. Assembly algorithm flowchart depicting automatic curation of text-based information into machine-readable data. Three example rows of data from the Ebola case study are shown for a single field (of 360 rows and 10 fields total). Trigger phrases are shown in purple and the numerical values of interest are shown in orange.

Figure 2. Assembly algorithm performance curves over time for the Ebola case study. Cumulative missingness is shown in orange, accuracy in teal, misidentification in purple, and data availability (at the source) in green.

205 **Funding Statement.** Research reported in this work was supported by the National
 206 Institutes of Health through an NIH Director’s New Innovator Award DP2-MD012722. The
 207 funder had no role in study design, data collection and analysis, decision to publish, or
 208 preparation of the manuscript.

209

210 **Conflicts of Interest.** The authors declare no conflicts of interest.

211

References.

1. Disease outbreaks by year. The World Health Organization. <https://www.who.int/csr/don/archive/year/en/>
2. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci.* 2001; 356(1411):983–9.
3. Zoonotic & infectious disease. Center for One Health Research. <https://deohs.washington.edu/cohr/zoonotic-infectious-disease>
4. Gollakner R, Capua I. Is COVID-19 the first pandemic that evolves into a panzootic? *Vet Ital.* 2020;56(1):7–8.
5. Greger M. The human/animal interface: emergence and resurgence of zoonotic infectious diseases. *Crit Rev Microbiol.* 2007;33(4):243–99.
6. Findlater A, Bogoch II. Human Mobility and the Global Spread of Infectious Diseases: A Focus on Air Travel. *Trends Parasitol.* 2018;34(9):772–83.
7. Dimala CA, Kadia BM, Nji MAM, Bechem NN. Factors associated with measles resurgence in the United States in the post-elimination era. *Sci Rep.* 2021;11(1):51.
8. Papachrisanthou MM, Davis RL. The Resurgence of Measles, Mumps, and Pertussis. *J Nurse Pract.* 2019;15(6):391–5.
9. Government of Samoa Twitter Account. November 22, 2019 (3:17 AM EST). <https://twitter.com/samoagovt/status/1197790948178051074>
10. Government of Samoa Twitter Account. December 8, 2019 (4:49 PM EST). <https://twitter.com/samoagovt/status/1203793768182235136>
11. Situation Épidémiologique, Lundi 6 août 2018. Ministère de la Santé République Démocratique du Congo. https://mailchi.mp/70213f4262fb/ebola_kivu_6aout/
12. Situation Épidémiologique, Mercredi 31 juillet 2019. Ministère de la Santé République Démocratique du Congo. https://mailchi.mp/sante.gouv.cd/ebola_kivu_31juil19/
13. Middle East respiratory syndrome coronavirus (MERS-COV) – Republic of Korea, 30 May 2015. The World Health Organization. <https://www.who.int/csr/don/30-may-2015-mers-korea/en/>
14. Middle East respiratory syndrome coronavirus (MERS-COV) – Republic of Korea, 9 June 2015. The World Health Organization. <https://www.who.int/csr/don/09-june-2015-mers-korea/en/>
15. George DB, Taylor W, Shaman J, et al. Technology to advance infectious disease forecasting for outbreak management. *Nat Commun.* 2019;10(1):3932.
16. Majumder MS, Santillana M, Mekaru SR, 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 Surveill.* 2016;2(1):e30.
17. Tuite AR, Fisman DN. The IDEA model: A single equation approach to the Ebola forecasting challenge. *Epidemics.* 2018;22:71–7.
18. Fisman DN, Hauck TS, Tuite AR, Greer AL. An IDEA for short term outbreak projection: nearcasting using the basic reproduction number. *PLoS One.* 2013;8(12):e83622.
19. Fisman D, Khoo E, Tuite A. Early Epidemic Dynamics of the West African 2014 Ebola Outbreak: Estimates Derived with a Simple Two-Parameter Model. *PLoS Curr.* 2014; doi: ecurrents.outbreaks.89c0d3783f36958d96ebbae97348d571.

20. Betti MI, Heffernan JM. A simple model for fitting mild, severe, and known cases during an epidemic with an application to the current SARS-CoV-2 pandemic. *Infect Dis Model.* 2021;6:313–23.
21. Greer AL, Spence K, Gardner E. Understanding the early dynamics of the 2014 porcine epidemic diarrhea virus (PEDV) outbreak in Ontario using the incidence decay and exponential adjustment (IDEA) model. *BMC Vet Res.* 2017;13(1):8.
22. Majumder MS, Kluberg S, Santillana M, et al. 2014 Ebola Outbreak: Media Events Track Changes in Observed Reproductive Number. *PLoS Curr.* 2015; doi: 10.1371/currents.outbreaks.e6659013c1d7f11bdab6a20705d1e865.
23. Price DJ, Shearer FM, Meehan MT, et al. Early analysis of the Australian COVID-19 epidemic. *Elife.* 2020;9:e58785.
24. Majumder MS, Cohn EL, Santillana M, Brownstein JS. Estimation of pneumonic plague transmission in Madagascar, August–November 2017. *PLoS Curr.* 2018; doi: ecurrents.outbreaks.1d0c9c5c01de69dfbfff4316d772954f.
25. Pan A, Liu L, Wang C, et al. Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. *JAMA.* 2020;323(19):1915–23.
26. Majumder MS, Cohn EL, Mekaru SR, et al. Substandard Vaccination Compliance and the 2015 Measles Outbreak. *JAMA Pediatr.* 2015;169(5):494–5.
27. Fisman D, Tuite A. Projected Impact of Vaccination Timing and Dose Availability on the Course of the 2014 West African Ebola Epidemic. *PLoS Curr.* 2014; doi: 10.1371/currents.outbreaks.06e00d0546ad426fed83ff24a1d4c4cc.
28. Majumder MS, Nguyen CN, Cohn EL, et al. Vaccine compliance and the 2016 Arkansas mumps outbreak. *Lancet Infect Dis.* 2017;17(4):361–2.
29. Zhao S, Stone L, Gao D, He D. Modelling the large-scale yellow fever outbreak in Luanda, Angola, and the impact of vaccination. *PLoS Negl Trop Dis.* 2018; 12(1):e0006158.
30. Majumder MS, Nguyen CM, Mekaru SR, Brownstein JS. Yellow fever vaccination coverage heterogeneities in Luanda province, Angola. *Lancet Infect Dis.* 2016;16(9):993–5.
31. Mizumoto K, Endo A, Chowell G, et al. Real-time characterization of risks of death associated with the Middle East respiratory syndrome (MERS) in the Republic of Korea, 2015. *BMC Med.* 2015;13:228.
32. Majumder MS, Kluberg SA, Mekaru SR, Brownstein JS. Mortality Risk Factors for Middle East Respiratory Syndrome Outbreak, South Korea, 2015. *Emerg Infect Dis.* 2015;21(11):2088–90.
33. Rahman A, Sarkar A. Risk Factors for Fatal Middle East Respiratory Syndrome Coronavirus Infections in Saudi Arabia: Analysis of the WHO Line List, 2013-2018. *Am J Public Health.* 2019;109(9):1288–93.
34. Fiebig L, Soyka J, Buda S, et al. Avian influenza A(H5N1) in humans: new insights from a line list of World Health Organization confirmed cases, September 2006 to August 2010. *Euro Surveill.* 2011;16(32):19941.
35. Yang Y, Hsu C, Lai C, et al. Impact of Comorbidity on Fatality Rate of Patients with Middle East Respiratory Syndrome. *Sci Rep.* 2017;7(1):11307.

36. Challen R, Brooks-Pollock E, Read JM, et al. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ*. 2021;372:n579.
37. Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis*. 2020;20(6):669–77.
38. Majumder MS, Brownstein JS, Finkelstein SN, et al. Nosocomial amplification of MERS-coronavirus in South Korea, 2015. *Trans R Soc Trop Med Hyg*. 2017;111(6):261–9.
39. Cowling BJ, Park M, Fang VJ, et al. Preliminary epidemiological assessment of MERS-CoV outbreak in South Korea, May to June 2015. *Euro Surveill*. 2015;20(25):7–13.
40. Majumder MS, Rivers C, Lofgren E, Fisman D. Estimation of MERS-Coronavirus Reproductive Number and Case Fatality Rate for the Spring 2014 Saudi Arabia Outbreak: Insights from Publicly Available Data. *PLoS Curr*. 2014; doi: ecurrents.outbreaks.98d2f8f3382d84f390736cd5f5fe133c.
41. Ogden NH, Fazil A, Safronetz D, et al. Risk of travel-related cases of Zika virus infection is predicted by transmission intensity in outbreak-affected countries. *Parasit Vectors*. 2017;10:41.
42. Majumder MS, Mandl KD. Early Transmissibility Assessment of a Novel Coronavirus in Wuhan, China. *SSRN*. First Posted: January 23, 2020; Last Updated: January 26, 2020.
43. Lourenco J, Monteiro ML, Valdez T, et al. Epidemiology of the Zika Virus Outbreak in the Cabo Verde Islands, West Africa. *PLoS Curr*. 2018; doi: ecurrents.outbreaks.19433b1e4d007451c691f138e1e67e8c.
44. White LF, Pagano M. Transmissibility of the Influenza Virus in the 1918 Pandemic. *PLoS One*. 2008;3(1):e1498.
45. Majumder MS, Mandl KD. Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. *Lancet Glob Health*. 2020; doi: S2214-109X(20)30113-3.
46. Chen IY, Pierson E, Rose S, et al. Ethical Machine Learning in Health Care. *arXiv*. First Posted: September 22, 2020; last Updated: October 8, 2020.

332 **Author Contributions.**

333 Maimuna S. Majumder, PhD: Study design; data acquisition, analysis, and interpretation;
334 drafting the work; critical revision of the work

335 Sherri Rose, PhD: Study design; data interpretation; critical revision of the work

336

337 **Address for Correspondence.**

338 Name: Maimuna S. Majumder

339 Email Address: maimuna.majumder@childrens.harvard.edu

340 Mailing Address: 300 Longwood Avenue, Boston, MA 02115