Data Challenges in Disease Response: The 2014 Ebola Outbreak and Beyond

KUSH R. VARSHNEY, DENNIS WEI, KARTHIKEYAN NATESAN RAMAMURTHY and ALEKSANDRA MOJSILOVIĆ, IBM Thomas J. Watson Research Center

Categories and Subject Descriptors: E.0 [**Data**]: General

General Terms: Algorithms, Legal Aspects, Documentation, Human Factors

Additional Key Words and Phrases: Disaster response, humanitarian crisis, data quality, open data

ACM Reference Format:
Kush R. Varshney, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Aleksandra Mojsilović. 2015. Data challenges in disease response: The 2014 Ebola outbreak and beyond. ACM J. Data Inf. Qual. 6, 2–3, Article 5 (June 2015), 3 pages.
DOI: http://dx.doi.org/10.1145/2742550

Data-driven decision making has much promise for effectively responding to large-scale disease outbreaks. Data can be used to track the status of the crisis on the ground, inform resource allocation and intervention decisions, feed into simulations for what-if decision support, illuminate causal factors of disease propagation, and engage the citizenry. However, in disease outbreak situations, we are confronted with several data collection and quality challenges [Raschid 2007]. An example of such an outbreak is Ebola virus disease (EVD) in West Africa in 2014. It was the most severe outbreak of EVD in history [WHO Ebola Response Team 2014]; at the end of October 2014, the World Health Organization (WHO) announced that nearly 14,000 cases and 5,000 deaths had been officially reported since the beginning of the outbreak. EVD resulted in a humanitarian crisis in West Africa not only through morbidity and mortality directly attributable to the disease but also due to second- and third-order effects such as diversion of resources and slowing of economic activity (R. Garfield, personal communication, October 2014).

In this article, taking EVD as our working example, we discuss a variety of data quality issues in disease response. We organize our discussion around three broad categories of data summarized in Table I: disease incidence, interventions and general humanitarian conditions, and mobility of individuals, along with associated challenges including underreporting and incompleteness, lack of a consolidated and properly cataloged data repository, and legal issues.

Disease incidence. The starting point for decision making is accurate data on disease incidence. Obtaining such data with sufficient temporal and spatial granularity can be a challenge. In the case of EVD, whereas national-level case and death statistics were

Authors' address: K. R. Varshney, D. Wei, K. N. Ramamurthy, and A. Mojsilović, IBM T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598.

DOI: http://dx.doi.org/10.1145/2742550

Table I. Categories of Data Sources

Category	Details/Examples	Challenges	Uses
Disease incidence	Counts of different types of cases and deaths	Underreporting, spatial and temporal granularity	Predict and plan for current and future cases
Interventions and general conditions	News reports, medical facility capacity and inventory, maps, population statistics, satellite imagery	Lack of unified repository, heterogeneity, redundancy, provenance and trust, availability, completeness, legal issues	Logistically support responses to the disease and many other humanitarian and economic challenges
Mobility	Call detail records	Privacy and security, government regulations	Trace contacts, predict future cases

widely reported, statistics at subnational levels (county/district/prefecture or below) were less available and often depended on the success or failure of a small number of local workers. A related and perhaps more troubling problem is underreporting of disease incidence. During the height of the EVD crisis, it was readily acknowledged by WHO that the official counts were significantly lower than the true incidence. For example, Meltzer et al. [2014] estimated 2.5 actual cases for every reported case. Underreporting is caused by the synergy of several factors: resistance from the populace to interact with dispatched health teams; patients hiding due to fear of authority; not enough disease surveillance officers and laboratory workers; and poor communication and information infrastructure [Belluz 2014]. Accordingly, both underreporting and spatiotemporal granularity can be improved through education and engagement of the citizenry [Chunara et al. 2013]; increased resources on the ground; better coordination of reporting by health ministries and aid organizations; and stronger information and communication systems, including the deployment of mobile apps and SMS-based data collection mechanisms. Statistical methods can be applied to estimate the true case incidence [Hook and Regal 1995].

Interventions and general conditions. Beyond incidence data, disease response can be aided by data sources related to interventions and general conditions, ranging from news reports to medical facility capacity and inventory to transportation maps to population statistics. Certain datasets that are incomplete or otherwise lacking in quality, such as maps and medical facility locations, may be enhanced using satellite image analysis by crowdsourcing or algorithms [Haklay and Weber 2008; Varshney et al. 2015]. Other datasets can be improved by similar means as for incidence data.

The preceding datasets are of various type, provenance, spatial and temporal resolution and coverage, and quality. Many are accessible on the Internet but are not unified into a common repository. To begin to address this challenge in the EVD case, on October 18, 2014, in New York City, in an event known as the Ebola Open Data Jam that attracted more than 100 volunteers, we led a group in cataloging as many data sources as possible. We asked volunteers to note metadata about the datasets (e.g., the items mentioned in the first sentence of this paragraph), possible analytics that they could enable, and their terms of use (datasets have varying terms with many meeting the criteria for *open data*). This activity uncovered approximately 50 heterogeneous data sources having the various quality challenges mentioned earlier.

We then published these data sources—some directly hosted and others federated—onto a cloud-based platform accessible at http://eboladata.org. The platform was built using DKAN, a Drupal-based open data platform developed by NuCivic, and allowed us to catalog the datasets and their metadata according to the World Wide Web Consortium's Data Catalog Vocabulary. The platform's application programming interface can support an ecosystem of data analyses and visualizations relevant to disease response, including mashups that integrate several data sources. Nevertheless, such integration has challenges due to heterogeneity and lack of record linkage across sources.

Redundancy and provenance are also issues, as some sources may reproduce or otherwise incorporate data from others. Some data sources we cataloged had terms that did not allow publishing; such a challenge may be addressed by contacting the owners and proposing alternative terms to which they may agree. Working with volunteers of different backgrounds and motivations resulted in varying levels of cataloging completeness and quality, which can be addressed in future events through more targeted volunteer selection and more initial training.

Mobility of individuals. Mobile telephone penetration in Africa is very high, and call detail records (CDRs) provide an excellent source of information on population movements that can be used to predict disease spread and plan responses [Wesolowski et al. 2014]. CDRs can also aid in *contact tracing*—the identification of people who have recently interacted with infected individuals, which is a crucial task in controlling outbreaks. The primary mobile carrier in West Africa has a history of releasing CDRs in such a way that individual privacy is preserved but inferences drawn from the data release are valid [de Montjoye et al. 2014]. However, the governments of the affected countries have not yet permitted the release of (anonymized) CDRs. This challenge may be addressed through lobbying of the appropriate government ministries.

Summary. We have identified several data and information quality challenges in outbreak response through data-driven means. By doing so, we have identified the need for unified platforms and toolkits that allow the collection, publishing, curation, federation, validation, and management of data that support different collection mechanisms, including mobile apps, SMS, and surveys, and can rapidly be deployed.

ACKNOWLEDGMENTS

The authors thank Steven Adler, Jayant Kalagnanam, Kala Fleming, Kun Hu, James Kaufman, Meenal Pore, Osamuyimen Stewart, and Jinfeng Yi of IBM, Andrew Hoppin of NuCivic, Richard Garfield of the Centers for Disease Control and Prevention, and John Brownstein of HealthMap for discussions and support.

REFERENCES

J. Belluz. 2014. No One Knows Exactly How Bad West Africa's Ebola Epidemic Is. Retrieved May 11, 2015, from http://www.vox.com/2014/10/6/6889037/reporting-ebola-epidemic-virus-outbreak.

R. Chunara, M. S. Smolinski, and J. S. Brownstein. 2013. Why we need crowdsourced data in infectious disease surveillance. *Current Infectious Disease Reports* 15, 4, 316–319.

Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. 2014. D4D-Senegal: The second mobile phone data for development challenge. arXiv:1407.4885.

M. Haklay and P. Weber. 2008. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing* 7, 4, 12–18.

E. B. Hook and R. R. Regal. 1995. Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews* 17, 2, 243–264.

M. I. Meltzer, C. Y. Atkins, S. Santibanez, B. Knust, B. W. Petersen, E. D. Ervin, S. T. Nichol, I. K. Damon, and M. L. Washington. 2014. Estimating the future number of cases in the Ebola epidemic—Liberia and Sierra Leone, 2014–2015. *Morbidity and Mortality Weekly Report* 63, Suppl. 3, 1–14.

L. Raschid. 2007. Information integration and dissemination for disaster data management. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines and Domains*. 333–335.

K. R. Varshney, G. H. Chen, B. Abelson, K. Nowocin, V. Sakhrani, L. Xu, and B. L. Spatocco. 2015. Targeting villages for rural development using satellite image analysis. *Big Data* 3, 1, 41–53.

A. Wesolowski, C. O. Buckee, L. Bengtsson, E. Wetter, X. Lu, and A. J. Tatem. 2014. Commentary: Containing the Ebola outbreak—the potential and challenge of mobile network data. *PLOS Currents Outbreaks* 1.

WHO Ebola Response Team. 2014. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *New England Journal of Medicine* 371, 16, 1481–1495.

Received October 2014; revised February 2015; accepted February 2015