The dynamics of Ebola underreporting during the 2014 West Africa epidemic

SV Scarpino

April 13, 2015

Introduction

Infectious disease surveillance data can be unreliable during unfolding crises in which resources are limited and public health authorities have poor access to affected communities. In the ongoing 2014 Ebola epidemic in West Africa, surveillance efforts primarily detect cases that are treated in healthcare facility, and likely miss the sizable fraction of cases that do not seek care in a clinical setting [1, 2]. The CDC recently used a factor of 2.5 to correct for this underreporting in Sierra Leone and LIberia, but acknowledged that this quantity is essentially unknown [1]. Accurate outbreak projections and assessments of intervention strategies depend on reliable estimates of underreporting rates. However, underreporting can be a very dynamic process, potentially varying in time, space, and/or with outbreak size, and driven by intrinsic properties of the pathogen, human behavior, diagnostic practices, and the healthcare infrastructure. Mechanistic models that capture these factors is critically important for estimating, reducing, and correctly accounting for underreporting, but they do not yet exist.

The primary difficulty we have faced in estimating underreporting is the limited data commonly available during an outbreak, namely confirmed and suspected cases and mortalities. From these data alone, one cannot estimate the rate of underreporting without making strong modeling assumptions and, even with such assumptions, we often lack statistical power to make precise estimates. However, new types of data are available from the ongoing EVD outbreak, including digital data - Facebook, Twitter, cell phones, HealthMap, etc - and Ebola virus genomic data [3]. These data, when coupled with an appropriate mathematical and statistical framework, provide a novel opportunity to model underreporting. For example, Facebook surveys can poll urban populations in West Africa regarding healthcare seeking behavior and, phylodynamic models can directly estimate the rate of underreporting from pathogen sequence data.

Here, we propose to construct an integrative framework, combining a mechanistic model of underreporting factors with survey-based and phylodynamic-based inference, to estimate dynamic rates of underreporting in the ongoing EVD outbreak. The results of this study will have an immediately impact, in enabling better EVD projections and analyses of intervention policies, and more broadly advance our understanding of and statistical analysis of reporting biases in emerging epidemics.

Phylodynamic approach

Underreporting can cause a mismatch between incidence estimated from case data and incidence reconstructed from genetic data. For example, if there is a constant level of underreporting, case count data will reflect lower transmission rates and lead to underestimation proportional to the underreporting rate. However, the extent of genetic variation among viral sequences taken from the same set of cases will refect the true, larger population size of circulating viruses, and lead to estimates closer to the true incidence. This is true even if viral sequences are only collected from reported individuals, assuming reported and un-reported individuals are mixing with each other. In preliminary analysis of EBOV genome sequences, we estimated that 58% (20–99%) of all cases were included in the sample. However, over 70% of confirmed patients in Sierra Leone were sequenced and included in the sample [3]. The discrepancy suggests that underreporting of cases may be between 0–70%, with the most likely value being 17%. The high degree of uncertainty is a result of inadequacies in available methods. We propose developing an integrative, mathematical framework for estimating underreporting. This approach will estimate an underreporting rate by jointly fitting epidemiological models to case and genomic data using Bayesian statistics.

Phylodynamic Methods

The primary tool for modern population genetic inference is coalescent theory, which provides a retrospective, mathematical framework for relating genetic variation to historical evolutionary processes [4]. Coalescent theory provides a framework to infer the evolutionary history of a population by sampling individuals in the present [4, 5, 6]. Consider a population in which individuals are related by a shared ancestry rooted at their Most Recent Common Ancestor (MRCA). Going forward in time and starting from the MRCA, the population diverges with lineages forming and dying. Looking backwards, lineages fuse, reducing in number until only a single lineage remains; the coalescent is a quantitative, probabilistic framework for determining when lineages join, or coalesce, backwards in time [4]. Because the coalescent considers neutral genetic variation, all pairs of existing lineages are equally likely to coalesce [4, 7, 8]. The result is a genealogy tracing the current individuals backwards in

time to the MRCA. The parameters of a coalescent model describe this stochastic, genealogical process. The rate that these lineages are born and die is also a function of the selective evolutionary forces acting on the population[4]. Therefore, selection, demography, and other evolutionary processes will leave signatures in the shape of genealogies [4, 5, 9, 10]. The expected coalescent time and the rate of coalescent are both highly sensitive to changes in ecological and evolutionary dynamics.

An Integrative Approach

Each data set, survey, genomic, and cases data, will contain information on the underreporting rate of EVD cases. We propose to jointly infer an underreporting rate for each country by fitting a single model of Ebola virus transmission using data from both sets. The underlying model will be based on a recently published continuous-time stochastic compartment model that partitions Ebola transmission among the community, health-care settings, and funerals [11]. By partitioning transmission into these three sectors, it may be possible to infer whether spatiotemporal variation in underreporting varies by transmission setting. For example, community transmission events may be more likely to result in unreported cases than hospital transmission events. To fit such a model to our datasets, we will apply Sequential Monte Carlo particle filtering - an approach that has been demonstrated as an effective statistical method for jointly fitting case and genomic data to a transmission model [12]. The results of this analysis will both have an immediate impact public health decision making during this EVD outbreak, result in a toolkit for use in future disease outbreaks, and improve our scientific understanding of how underreporting effects the population genomic variation of viral pathogens.

References

[1] Martin I Meltzer, Charisma Y Atkins, Scott Santibanez, Barbara Knust, Brett W Petersen, Elizabeth D Ervin, Stuart T Nichol, Inger K Damon, and Michael L Washington. Estimating the future number of cases in the ebola epidemic—liberia and sierra leone, 2014–2015. *MMWR Surveill Summ*, 63(suppl 3):1–14, 2014.

[2] WHO Ebola Response Team. 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, 2014.

[3] Stephen K Gire, Augustine Goba, Kristian G Andersen, Rachel SG Sealfon, Daniel J Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, et al. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[4] John Wakeley. *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers Greenwood Village, Colorado, 2009.

[5] Alexei J Drummond, Andrew Rambaut, Beth Shapiro, and Oliver G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192, 2005.

[6] J Wakeley. Metapopulations and coalescent theory. *Ecology, genetics and evolution of metapopulations Hanski I, Gaggiotti OE*, pages 175–198, 2004.

[7] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[8] John FC Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

[9] John Parsch, Colin D Meiklejohn, and Daniel L Hartl. Patterns of dna sequence variation suggest the recent action of positive selection in the janus-ocnus region of drosophila simulans. *Genetics*, 159(2):647–657, 2001.

[10] Masatoshi Nei and Naoyuki Takahata. Effective population size, genetic diversity, and coalescence time in subdivided populations. *Journal of Molecular Evolution*, 37(3):240–244, 1993.

[11] Abhishek Pandey, Katherine E Atkins, Jan Medlock, Natasha Wenzel, Jeffrey P Townsend, James E Childs, Tolbert G Nyenswah, Martial L Ndeffo-Mbah, and Alison P Galvani. Strategies for containing ebola in west africa. *Science*, page 1260612, 2014.

[12] David A Rasmussen, Oliver Ratmann, and Katia Koelle. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS computational biology*, 7(8):e1002136, 2011.