Leveraging pathogen sequence and contact tracing data to enhance vaccine trials in emerging epidemics

Rebecca Kahn¹, Rui Wang^{2,3}, Sarah V. Leavitt⁴, William P. Hanage¹, Marc Lipsitch^{1,5}

¹ Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA

² Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, USA

³ Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA

⁴ Department of Biostatistics, School of Public Health, Boston University, Boston, Massachusetts, USA

⁵ Department of Immunology and Infectious Diseases, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA

Abstract

Introduction: Advance planning of the design and analysis of vaccine trials conducted during infectious disease outbreaks increases our ability to rapidly define the efficacy and potential impact of a vaccine and inform public health response. Vaccine efficacy against infectiousness (VE_i) is an important measure for understanding the full impact of a vaccine, yet it is currently not identifiable in many vaccine trial designs because it requires knowledge of the vaccination status of infectors. Recent advances in pathogen genomics have improved our ability to accurately reconstruct transmission networks. We aim to assess if augmenting classical randomized controlled trial designs with pathogen sequence and contact tracing data can permit these trials to estimate VE_i .

Methods: We develop a transmission model with a vaccine trial in an outbreak setting, incorporate pathogen sequence evolution data and sampling as well as contact tracing data, and assign probabilities to likely infectors. We then propose and evaluate the performance of an estimator of VE_i .

Results: We find that under perfect knowledge of infector-infectee pairs, we are able to accurately estimate VE_i . Use of sequence data results in imperfect reconstruction of the transmission networks, biasing estimates of VE_i towards the null, with approaches using deep sequence data performing better than approaches using consensus sequence data. Inclusion of contact tracing data reduces the bias.

Conclusion: Pathogen genomics enhance identifiability of VE_i from individually randomized controlled trials, but imperfect transmission network reconstruction biases the estimates towards the null and limits our ability to detect VE_i . Given the consistent direction of the bias, estimates obtained from trials using these methods will provide lower bounds on the true VE_i . A combination of sequence and epidemiologic data results in the most accurate estimates, underscoring the importance of contact tracing in reconstructing transmission networks.

Introduction

Vaccine trials conducted during epidemics of emerging infectious diseases provide an important opportunity to test the safety and efficacy of vaccine candidates. Increasing our ability to quickly and accurately understand the impact of a vaccine candidate in the urgent setting of an outbreak is critical for enhancing public health response. The use of the ring vaccination strategy in the *Ebola ça Suffit* trial during the 2013-2016 West African Ebola outbreak highlighted the importance of developing innovative designs for trials conducted during an ongoing outbreak.¹ It also underscored the need to think through trial design and analysis strategies in advance in order to expedite the rollout of a vaccine trial once an outbreak starts and to identify the best methods for obtaining high quality efficacy estimates in outbreak settings.²

Multiple components of vaccine efficacy can be estimated from a vaccine trial.³ Individually randomized controlled trials (iRCTs) estimate vaccine efficacy against susceptibility to infection (VE_s), the direct effect of the vaccine on vaccinated individuals.³ If reducing susceptibility to infection is the only effect of the vaccine, then this measure, combined with information on contact network structure and pathogen transmission dynamics, can be used to estimate the total effect of a vaccination program, a combination of the direct and indirect (i.e. herd immunity) effects. Vaccine efficacy against infectiousness (VE_i), the reduction in onward transmission from a vaccinated person who is infected compared to an unvaccinated infected person, is another important measure for understanding the impact of a vaccine.³ Even if a vaccine does not protect everyone who is vaccinated from getting infected, its impact on infectiousness for those who are vaccinated but nevertheless become infected plays a critical role in both outbreak dynamics and also cost-effectiveness of a vaccine program. The significance of understanding interventions' effects on future transmission is exemplified by the efforts of HIV treatment-as-prevention programs to reduce patients' viral loads to undetectable levels in order to prevent onward transmission.^{4,5}

In order to estimate VE_i , the vaccination status of infectors must be known. VE_i is therefore potentially measurable in household studies^{6,7} and partner transmission studies, such as HIV vaccine trials⁸ because in these settings, infector-infectee pairs can be identified (by assuming that household members or partners are the infectors), and thus the vaccination status of infectors is known. However, VE_i is not currently identifiable in population-level vaccine trials, such as those often conducted during an infectious disease outbreak, because the transmission network, and consequently the vaccination status of infectors, are typically unknown.

Recent advances in pathogen genomics have improved our ability to accurately reconstruct transmission networks.^{9–15} The West African Ebola epidemic and the ongoing COVID-19 pandemic have demonstrated our growing capacity to use sequence data in outbreak settings,^{16–21} and recent work has highlighted the potential for deep sequence data to add resolution to transmission networks.^{22–24} We aim to assess if augmenting classical randomized controlled trial designs with pathogen sequence data, as well as contact tracing data, would

permit these trials to estimate VE_I by reconstructing transmission networks and identifying the trial status of infectors.

Methods

We define θ as the risk ratio for becoming infected if one receives vaccine vs. control, or $1 - VE_S$, and Φ as the relative infectiousness of a vaccinated person who is infected compared to a control who is infected, or $1 - VE_I$. At the conclusion of a vaccine trial, the ratio of the proportion of people infected by vaccinated individuals to the proportion of people infected by controls is a product of both the vaccine's effect on susceptibility to infection and its effect on infectiousness among those who are infected. With knowledge of who infected whom, using the ratio of infector vaccination status, we can therefore calculate VE_I :

$$\theta\Phi = (1 - VE_S)(1 - VE_I) = \frac{\# \text{ infected by vacc} / \# \text{ vacc}}{\# \text{ infected by control} / \# \text{ control}}$$

$$VE_I = 1 - \frac{\# \text{ infected by vacc} / \# \text{ vacc}}{\# \text{ infected by control} / \# \text{ control}} / (1 - VE_S)$$

We simulate a compartmental network model of an outbreak, together with a vaccine trial, the details of which have been previously described.²⁵ Individuals are grouped into communities, with many connections between individuals in the same community and fewer between individuals in different communities. Introduction of infection into the network occurs at a time-varying rate, and the disease natural history in the communities follows a stochastic susceptible, exposed, infectious, recovered (SEIR) model, with Ebola-like parameters (Table 1). Each individual has a daily probability of infection from their infectious contacts in the network. Individuals are enrolled into an iRCT, with 50% randomized to vaccine and 50% to control. The vaccine's efficacy against susceptibility to infection is "leaky", with 60% efficacy ($VE_S = 0.60$), meaning upon each exposure, the vaccine reduces a vaccinated individual's chance of infection by 60%. The vaccine's efficacy against infectiousness is 30%, meaning infectiousness among infected vaccinated individuals is 30% lower than among infected unvaccinated individuals ($VE_I = 0.30$). Table 2 shows the number of infections expected for each type of infector-infectee pair from the trial simulations.

To estimate VE_I , we first make the unrealistic assumption of complete knowledge of the transmission network, with perfect ascertainment of who infected whom and their infection and recovery times. We then relax the assumption of perfect knowledge of who infected whom. Using the R package *seedy*,²⁶ we incorporate pathogen evolution and sampling of both consensus and deep sequence data into the simulations, specifying parameters such as genome length, mutation rate, and bottleneck size (Table 1). As the choice of parameters, particularly mutation rate and bottleneck size, greatly impacts our ability to reconstruct transmission

networks,²² we vary parameters across simulations to assess their impact on our ability to estimate VE_i .

For each infectee we then assign a probability to each potential source of infection, based on comparisons of the sequence data for the candidate source(s) and each index case using four different approaches. In the first two approaches we use consensus sequence data. First, we assign probabilities to potential infectors based on the inverse of the genetic distance between the infectee and potential infectors. Second, we use a geometric-Poisson approximation of SNP distance to assign probabilities to potential infectors; this approach assumes genetically similar sequences are more likely to be infector-infectee pairs, while also accounting for mutation rate and times of infection.²⁶ Third, we weight potential infectors by the number of rare variants (i.e. minority variants not seen in the consensus sequence that are rare in the population) they share with each infectee, which may be identified through deep sequence data and has previously been shown to provide additional resolution.²² Fourth, we combine the second and third approaches, using the consensus sequence data in the event that no shared minority variants for an infectee are identified through deep sequence data.²² For all four approaches, we then weight the probabilities identified through the sequence data by the probability of infection given the time of symptom onset of the infectee and potential infector(s) based on the serial interval distribution (i.e. the time between when an infector becomes symptomatic and their infectee becomes symptomatic).

Using each of these approaches, we then estimate the ratio of the number of cases infected by a vaccinated person to the number of cases infected by a control. We do this in three ways for each approach (see supplemental text 1 for more details). First, we weight each identified potential infector by the probability assigned to them and sum the probabilities by vaccination status. Second, we split the probabilities for each infectee into clusters based on the largest gap in probabilities between potential infectors.²⁸ If the gap is larger than the specified threshold, we use the normalized probabilities from the infector(s) in the top cluster; otherwise we exclude that infectee from the analysis. Third, we use only the vaccination status of the most likely infector(s) for each infectee. Using the estimated ratio of the trial status of the potential infectors and the estimate of VE_s from the trial, we then estimate VE_i , using the equation above. To incorporate the data from the network obtained through contact tracing efforts during an epidemic, we also conduct all of the approaches described above in a data set restricted to only potential infectors who are contacts of the infectees (i.e. connections in the network model).

We propose the following procedure for estimating the standard errors of the estimates under the approaches that perform best. For a given simulation and estimate of VE_i , we obtain a bootstrap estimate of the standard error as follows. We first sample with replacement from the infected individuals. We then construct a bootstrapped data set using each infected individual from the sample and all of their potential infectors identified by the approach. We estimate VE_i from the bootstrapped data set and then repeat these steps 100 times. The standard deviation of the 100 bootstrapped estimates is the standard error of the VE_i estimate. This approach could be used with real data observed in a real trial and resembles the bootstrapping clusters

approach (i.e., clusters are treated as units for resampling) for clustered data.²⁹ Code is available: https://github.com/rek160/ADAGIO_WPA.

Results

As expected, under perfect knowledge of the transmission network, VE_i is estimated correctly (median of 500 simulations: estimate = 0.29, standard error = 0.19), while imperfect reconstruction of the transmission networks using sequence data results in bias towards the null away from the true VE_i of 0.30 (Figure 1). This imperfect reconstruction is due to the identification of multiple potential infectors for each infectee. For example, another infectee infected by the infector of an index case may share the same number of rare variants as the index case and thus be identified in the top cluster of potential infectors. Of the methods using only sequence data, the shared variant approach using deep sequence data and the hybrid approach return results closest to the true value of VE_i , while the approaches using consensus sequence data alone return estimates closer to the null. The approaches using clustering result in more accurate estimates of VE_i (Figure 1) than the methods weighting all possible infectors, or methods using only the most likely infector(s) (Figure S1).

In reality, sequence data are unlikely to be used in isolation, and adding epidemiologic data from the contact network decreases the bias. Using the infector(s) identified from the hybrid and shared variant approaches among potential infectors restricted to contacts results in median estimates of 0.28, close to the true value of 0.30 (Figures 1 & S1).

The ability to accurately reconstruct transmission networks was previously found to be influenced by parameters such as the bottleneck size and the mutation rate.²² Varying these and other parameters in our simulations show similar results to the baseline scenario (Figures S2-S5), with the hybrid and shared variant approaches performing worse with a lower mutation rate (Figure S2) and a lower bottleneck size (Figure S3), as expected because less shared variant information is available in these settings.²²

Discussion

In the case of an outbreak of an emerging infectious disease, the ability to rapidly define the efficacy and potential impact of a vaccine is crucial for improving public health and informing policy decisions. An important component of vaccine efficacy which is often overlooked is its ability not only to guard against acquisition of infection by vaccinated individuals, but also to prevent onward transmission from those who are vaccinated that nevertheless become infected. VE_i is important for fully understanding and modeling the impact of a vaccine and both sequence and contact tracing data have the potential to allow us to estimate VE_i in large individually-randomized controlled trials conducted during an epidemic. Previously, this estimate was only attainable from household and partner studies.⁶⁻⁸ Advance planning and understanding of the data requirements necessary are critical for obtaining efficacy estimates during the uncertain and urgent setting of an outbreak.

We find that while sequence and contact tracing data have the potential for enabling estimation of VE_i , misclassification of the trial status of infectors due to imperfect reconstruction of the transmission network leads to bias towards the null of VE_i estimates and overall limits our ability to detect an effect of the vaccine on infectiousness. Given the consistent direction of the bias, if an estimate is obtained in a trial using the methods described here, it is expected to be an underestimate of the true VE_i . The approaches using the top cluster of most likely infector(s) identified from the deep sequence shared variants and hybrid data perform the best of all of the methods using sequence data alone and remain the most accurate method when contact tracing data are incorporated. If deep sequencing data are not available, relying on contact tracing data becomes even more important. The substantial improvement in the estimates when restricting to contacts further underscores the importance of contact tracing for reconstructing transmission networks.

Previous work has pointed to the potential of shared variants identified in deep sequence data, to inform transmission.^{22–24} The intuition of this approach is that the pathogen population within an infected host is not composed of identical genomes, but contains some polymorphisms (depending on the population size and the mutation rate). If the transmission bottleneck is sufficiently large, more than one of these genotypes may be transmitted, and the finding that individuals share the resulting polymorphism is then a likely indication of transmission. The methods described here will therefore have variable efficacy for different pathogens. For example, influenza has a high mutation rate,⁹ so there is likely sufficient phylogenetic signal and within host variation to support reconstruction of the transmission network and estimation of VE_i . Initial genomics analyses of SARS-CoV-2 found a low mutation rate;³⁰ recently, however, there is evidence of minority variants detectable by deep sequencing,³¹ suggesting deep sequencing approaches have the potential to be used in ongoing vaccine trials to estimate VE_i .

Many simplifying assumptions have been made, which could be relaxed in future work. We assume perfect knowledge of infection and recovery times, allowing us to accurately identify the direction of transmission in infector-infectee pairs; in reality, particularly for pathogens with short incubation periods, the direction of transmission may be less clear. We also assume complete and correct sampling of sequence data (which in turn means that everyone in the community is a participant in the trial, as we assume, or at a minimum is followed up in the trial), full knowledge of the contact network, and complete contact tracing. Approaches such as those in the *TransPhylo* R package could be used to assess where cases are likely missing and the overall proportion of the outbreak that has been sampled.³² A naïve Bayes approach using additional data on individuals in the trial, such as demographic or geographic covariates, has been shown to improve reconstruction of the transmission network when limited sequence and/or contact tracing data are available.³³ Our methods further absorb the limitations of the *seedy* package, which assumes neutral evolution and does not permit superinfection, although this latter limitation is likely more of a concern for endemic rather than epidemic disease models.

Despite these simplifying assumptions, this work highlights the potential for existing data sources to be used in the midst of an outbreak to estimate a key measure of vaccine efficacy. It further identifies the data sources that will lead to the most accurate estimation and can thus be used for better targeting of the limited resources available for data collection in the midst of an epidemic.

Acknowledgments

This research is funded by the Department of Health and Social Care using UK Aid funding and is managed by the NIHR. RW was supported by funding from R01 AI136947 from the National Institute of Allergy and Infectious Diseases. SVL was supported by NIHGMS R01GM122876 and T32GM074905. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health and SocialCare or the NIH.

We thank Dr. Laura F. White for helpful discussion.

Conflicts of interest

ML discloses honoraria/consulting from Merck, Affinivax, Sanofi-Pasteur, and Antigen Discovery; research funding (institutional) from Pfizer, and an unpaid scientific advice to Janssen, Astra-Zeneca, and Covaxx (United Biomedical).

Figure 1.

The median VE_i estimates from 500 simulations with the baseline parameters, with a true VE_i of 0.3. “None” refers to simulations that use only sequence data, without incorporating any epidemiologic information from the network. “Restrict to contacts only” restricts the analysis of sequence data to potential infectors who are contacts from the network.

Table 1

Parameter	Value in baseline model	Values in supplement
R_0^{34}	1.5	
Incubation period ³⁵	9.7 days	
Infectious period ³⁵	5 days	
VE_s	0.6	0.8
VE_i	0.3	0.7
Number of communities	2	
Size of community	5,000	
Probability of connection within community	0.02	
Probability of connection between communities	0.001	
Importations from main population over trial period ²⁵	20	
Trial length (days)	300	
Genome length ⁹	18,958	
Mutation rate (per genome per generation)	0.012	0.003
Bottleneck (size of pathogen inoculum at transmission)	10	2
Cluster threshold ²⁸	0.2	0.1

Table 2

<i>Infecter (column)</i> <i>Infectee (row)</i>	Vaccinee	Control	Any participant	Ratio of infectees
Vaccinee	$apq\theta^2\Phi$	$apq\theta$	$apq\theta(1+\theta\Phi)$	θ
Control	$apq\theta\Phi$	apq	$apq(1+\Phi)$	
Any participant	$apq\theta\Phi(1+\theta)$	$apq(1+\theta)$	$apq(1+\theta)(1+\theta\Phi)$	
Ratio of infectors	$\theta\Phi$			

A proportion p is randomized to vaccine and to control. In the absence of vaccination, a proportion a would become infected, and a proportion $2q$ of all exposures to infection of participants would come from other trial participants (with $1-2q$ external exposures). $\theta = 1 - VE_s$, or the risk ratio for becoming infected if one receives vaccine vs. control, and $\Phi = 1 - VE_i$, or the relative infectiousness of a vaccinated person who becomes infected to a control who becomes infected.

Supplement text 1

Aim

In order to estimate VE_I , we need to estimate the **ratio** of the # people infected by a vaccinated person to the # infected by a control

$$VE_I = 1 - \frac{\frac{\# \text{ infected by vacc}}{\# \text{ vacc}}}{\frac{\# \text{ infected by control}}{\# \text{ control}}} / (1 - VE_S)$$

Because # vacc = # cont:

$$VE_I = 1 - \frac{\# \text{ infected by vacc}}{\# \text{ infected by control}} / (1 - VE_S)$$

Analysis

The ratio of the # infected by vacc / # infected by control comes from summing probabilities based on the vaccination status of potential infectors across all infectees.

Example for 1 infected person (probability could be obtained from any of the approaches described in the Methods)

Potential Infector	Trial Status	Probability	Cluster
A	Vacc	0.05	2
B	Vacc	0.40	1
C	Control	0.35	1
D	Control	0.10	2
E	Control	0.10	2

All prob: include all potential infectors and count each as their probability. We add 0.45 [0.05+0.4] to the numerator of the ratio above and 0.55 [0.35 + 0.10 + 0.10] to the denominator.

Cluster: Divide infectors into clusters based on biggest gap in probability (here: 0.35-0.10 = 0.25). Include top cluster if gap is greater than or equal to the threshold (in simulations above, threshold=0.2 so 0.25 meets this criteria). We add 0.40/(0.40+0.35) to the numerator of the ratio above and 0.35/(0.40+0.35) to the denominator because only 0.40 and 0.35 fall into the top cluster.

Max: Include only potential infector(s) that have the highest probability and count as 1. Here, we would add 1 to the numerator of the ratio and nothing to the denominator because 0.4, the maximum probability, is a vaccinated person. If two are tied for most likely then each counts ½.

Supplement

Figure S1. All methods

The median VE_i estimates from 500 simulations with the baseline parameters, with a true VE_i of 0.3. All approaches, including multiple clustering thresholds and the approach using the vaccination status of only the most likely infector(s) ("Max"), are shown, as well as the average standard errors for the best performing approaches.

*Cluster: 0.2 threshold; Cluster 1: 0.1 threshold used

Figure S2. Mutation rate = 0.003

The median VE_i estimates from 500 simulations with a lower mutation rate of 0.003, with a true VE_i of 0.3.

Figure S3. Bottleneck = 2

The median VE_i estimates from 500 simulations with a lower bottleneck size of 2, with a true VE_i of 0.3.

Figure S4. $VE_i = 0.7$

The median VE_i estimates from 500 simulations, with a true VE_i of 0.7.

Figure S5. $VE_S = 0.8$

The median VE_I estimates from 500 simulations with a VE_S of 0.8, with a true VE_I of 0.3.

References

1. Henao-Restrepo AM, Camacho A, Longini IM, Watson CH, Edmunds WJ, Egger M, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *Lancet*. 2017;389(10068):505–18.
2. Kahn R, Rid A, Smith PG, Eyal N, Lipsitch M. Choices in vaccine trial design in epidemics of emerging infections. *PLOS Med* [Internet]. 2018 Aug 7;15(8):e1002632. Available from: <https://doi.org/10.1371/journal.pmed.1002632>
3. Halloran ME, Longini IM, Struchiner CJ. Design and Analysis of Vaccine Studies: Introduction. Vol. 36. Springer; 2009.
4. Eisinger RW, Dieffenbach CW, Fauci AS. HIV viral load and transmissibility of HIV infection: undetectable equals untransmittable. *Jama*. 2019;321(5):451–2.
5. Wertheim JO, Kosakovsky Pond SL, Little SJ, De Gruttola V. Using HIV Transmission Networks to Investigate Community Effects in HIV Prevention Trials. *PLoS One*. 2011;6(11):1–7.
6. Datta S, Halloran ME, Longini Jr IM. Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual versus household. *Biometrics*. 1999;55(3):792–8.
7. Préziosi M-P, Halloran ME. Effects of pertussis vaccination on transmission: vaccine efficacy for infectiousness. *Vaccine*. 2003;21(17–18):1853–61.
8. Longini IMJ, Datta S, Halloran ME. Measuring Vaccine Efficacy for Both Susceptibility to Infection and Reduction in Infectiousness for Prophylactic HIV-1 Vaccines. *JAIDS J Acquir Immune Defic Syndr* [Internet]. 1996;13(5). Available from: https://journals.lww.com/jaids/Fulltext/1996/12150/Measuring_Vaccine_Efficacy_for_Both_Susceptibility.7.aspx
9. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? *PLoS Pathog*. 2018;14(2):e1006885.
10. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput Biol*. 2019;15(3):e1006930.
11. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol Biol Evol*. 2017;34(4):997–1007.
12. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity (Edinb)*. 2011;106(2):383.
13. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol*. 2014;10(1):e1003457.
14. Kenah E, Britton T, Halloran ME, Longini Jr IM. Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. *PLoS Comput Biol*. 2016;12(4):e1004869.
15. Kraemer MUG, Cummings DAT, Funk S, Reiner RC, Faria NR, Pybus OG, et al. Reconstruction and prediction of viral disease epidemics. *Epidemiol Infect* [Internet]. 2018/11/05. :1–7. Available from:

- <https://www.cambridge.org/core/article/reconstruction-and-prediction-of-viral-disease-epidemics/6CC411ED2ED7A896FCFF27F51B28FFCD>
16. Emmett KJ, Lee A, Khiabani H, Rabadan R. High-resolution genomic surveillance of 2014 ebolavirus using shared subclonal variants. *PLoS Curr*. 2015;7.
17. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* (80-). 2014;345(6202):1369–72.
18. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell*. 2015;161(7):1516–26.
19. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol*. 2004;160(6):509–16.
20. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* (80-) [Internet]. 2020 Jul 17;369(6501):297 LP – 301. Available from: <http://science.sciencemag.org/content/369/6501/297.abstract>
21. Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires caution. *Nat Microbiol* [Internet]. 2020;5(7):876–7. Available from: <https://doi.org/10.1038/s41564-020-0738-5>
22. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. *Am J Epidemiol* [Internet]. 2017 Nov 15;186(10):1209–16. Available from: <http://dx.doi.org/10.1093/aje/kwx182>
23. Martin MA, Lee RS, Cowley LA, Gardy JL, Hanage WP. Within-host *Mycobacterium tuberculosis* diversity and its utility for inferences of transmission. *Microb genomics* [Internet]. 2018/10/11. 2018 Oct;4(10):e000217. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/30303479>
24. Lee RS, Proulx J-F, McIntosh F, Behr MA, Hanage WP. Previously undetected super-spreading of *Mycobacterium tuberculosis* revealed by deep sequencing. *Elife*. 2020;9:e53245.
25. Kahn R, Hitchings M, Bellan S, Lipsitch M. Impact of stochastically generated heterogeneity in hazard rates on individually randomized vaccine efficacy trials. *Clin Trials*. 2018;15(2):207–11.
26. Worby CJ, Read TD. 'SEEDY'(Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens. *PLoS One*. 2015;10(6):e0129745.
27. Worby CJ, Chang H-H, Hanage WP, Lipsitch M. The distribution of pairwise genetic distances: a tool for investigating disease transmission. *Genetics*. 2014;198(4):1395–404.
28. Leavitt S V, Jenkins HE, Sebastiani P, Lee RS, Horsburgh Jr C, Tibbs A, et al. Estimation of the generation interval using pairwise relative transmission probabilities. *Biostatistics*. 2020;revise and resubmit.
29. Field CA, Welsh AH. Bootstrapping clustered data. *J R Stat Soc Ser B (Statistical Methodol)*. 2007;69(3):369–90.
30. Kupferschmidt K. Mutations can reveal how the coronavirus moves—but they're easy to overinterpret. *Science* (80-). 2020;

31. Jary A, Leducq V, Malet I, Marot S, Klement-Frutos E, Teyssou E, et al. Evolution of viral quasispecies during SARS-CoV-2 infection. *Clin Microbiol Infect* [Internet]. 2020 Jul 24;S1198-743X(20)30440-7. Available from: <https://pubmed.ncbi.nlm.nih.gov/32717416>
32. Xu Y, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M, et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. *PLoS Med* [Internet]. 2019 Oct 31;16(10):e1002961–e1002961. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/31671150>
33. Leavitt S V, Lee RS, Sebastiani P, Horsburgh Jr, CR, Jenkins HE, White LF. Estimating the relative probability of direct transmission between infectious disease patients. *Int J Epidemiol* [Internet]. 2020 Mar 24;49(3):764–75. Available from: <https://doi.org/10.1093/ije/dyaa031>
34. Althaus CL. Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa. *Public Libr Sci Curr Outbreaks*. 2014;1–9.
35. Hitchings MDT, Lipsitch M, Wang R, Bellan SE. Competing Effects Of Indirect Protection And Clustering On The Power Of Cluster-Randomized Controlled Vaccine Trials. *Am J Epidemiol*. 2018 Mar 7;