From: <mlipsitch@gmail.com> on behalf of Marc Lipsitch <mlipsitc@hsph.harvard.edu>

Position #1: https://academicpositions.harvard.edu/postings/8296 (this is the one I think would be of specific interest to biostatisticians).

Position #2: https://academicpositions.harvard.edu/postings/8298

The science of identifying who infected whom using densely sampled genomic data has progressed rapidly in recent years, from "best guess" heuristic approaches based on maximum similarity to more sophisticated probabilistic approaches (4), taking into account details of the transmission process (5,6) and more recently leveraging deep sequence data to provide considerably greater resolution (7–9). Likewise, choices of trial designs in infectious disease emergencies face tradeoffs of logistics/feasibility, sample size, and, some would argue, ethics (10–12). Cluster-randomised trials have some clear advantages, but one limitation of, for example, the ring vaccination design employed in Guinea for Ebola in 2014-5 is that the effect the trial measures is a complicated and context-specific combination of direct and indirect effects (10,13). We hypothesise that employing sequence data could considerably improve the value of cluster-randomised trials to measure and distinguish different measures of vaccine effectiveness. This approach is foreshadowed by HIV prevention trial designs and household randomisation trial designs that measure both VES and VEI (14,15). Enhancing trial designs

with sequence data is a promising way forward, but raises many questions of methodology and sampling strategy that we will address.

We aim to answer the research question: can augmenting classical RCT designs with pathogen sequence data permit these trials to estimate quantities not identifiable at present, viz. VE_I in iRCTs and VE_I and VE_S in cRCTs?

Objectives:

A1. Develop simulations of transmission in plausible outbreak settings for two example pathogens with different transmission modes, in which we track the infector of each case. Within each outbreak, we will simulate a range of trials with different designs (iRCT, traditional, ring and stepped-wedge cRCTs).

A2. Propose and test estimators for each quantity noted above using the augmented data available from the simulations to define circumstances in which these quantities are identifiable.

A3. Modify the simulations to incorporate pathogen sequence evolution data and sampling at the time of case ascertainment, and assign probabilities to each potential infector as the source for each case.

A4. Using these probabilistic identifications of sources, repeat the estimation of the novel quantities to determine under what circumstances sequence data can enable their reliable estimation.

A5. Determine optimal intensity and timing of pathogen sampling to maximize precision subject to resource constraints.