Protocol: Measuring the Ability of Disease Surveillance Strategies to Identify Transmission Links with Whole-Genome Sequencing Data

***Principal Investigator:***

Daniel E. Ho, JD, PhD1

Stanford University

Stanford, CA, USA

Email: deho@stanford.edu

Coauthors: Benjamin Anderson, MSc1; Derek Ouyang, MSc1; Alexis D’Agostino, MPP2; Brandon Bonin, DrPH2; Vit Kraushaar, MD2; Sarah L. Rudman, MD, MPH2

1 Stanford University, Stanford, California, USA

2 County of Santa Clara Public Health Department, San Jose, California, USA

3 California Department of Public Health, Sacramento, California

***Background:***

In the face of scarce public health resources, it is critical to understand which disease surveillance strategies provide reliable transmission information, what underutilized strategies may exist, and how strategies may serve as complements or substitutes.

***Aim:***

We use viral COVID-19 genomic data to assess the effectiveness of existing and potential strategies in identifying related cases.

***Study Design:***

We obtain COVID-19 patient case records from May 1 to December 31, 2021, from Santa Clara County’s (SCC) contact tracing database (CalCONNECT), which comprises deduplicated person records and incidents for each episode of COVID-19 disease and/or identification as a contact to COVID-19. The case records are pre-populated with patient name, date of birth, and home address from the laboratory which conducted the polymerase chain reaction (PCR) test, and subsequently augmented with other patient-level attributes through an automated survey and/or phone interview with a contact tracer.

We use information in CalCONNECT to categorize pairs of cases into disease surveillance strategies of interest. For *contact tracing*, close contacts reported by a positive case may be subsequently contacted by the contact tracer and encouraged to quarantine and test; if they end up testing positive, we can link the two positive individuals, and consider this link a unique disease surveillance artifact of contact tracing. We observe the artifacts of *mandated reporting* and associated outbreak investigation as exposure events categorized by location type and link all pairs within an event. We evaluate *schools*, *workplaces*, *jails*, and *long-term care facilities* (LTCFs) as distinct types of reporting. We also observe passive retrieval of links from other information gathered, such as a shared *home address* (collected upstream via testing) or shared *employer*, *school*, or nonresidential *location history* (collected via survey or interview). Altogether, these data allow us to identify proposed links between cases for an array of different strategies. We use these proposed links to measure the effectiveness of each disease surveillance strategy at identifying plausible COVID-19 transmission links. While we assess the informational value of each strategy independently, we also assess set differences of two strategies, such as removing same address links from the set of contact tracing links to assess inter-household spread.

We use all available, high-quality genomic sequences from PCR tests from SCC’s laboratory, as well as sequences from the Global Initiative on Sharing Avian Influenza Data (GISAID) COVID-19 database. We link sequences to cases in CalCONNECT from May 1 to December 31, 2021, resulting in a subset of cases matching a high-quality sequence.

***Outcomes:***

We define the informational value of each disease surveillance strategy as the fraction of sequenced links surfaced by the strategy that are biologically plausible according to genomic sequence data. Because not all pairs of cases are sequenced, we can only estimate informational value based on the sample of sequenced cases. We restrict the analysis to proposed pairs of cases whose episode dates are within 14 days of one another.

***Sample Size:***

The study’s timeline is based on a period of high sequencing rate in SCC. Our sample size is restricted to the availability of high-quality genomic sequences.

***Analysis Plan:***

To determine whether the genomic data support a proposed transmission link, we measure the genomic distance between sequenced pairs of cases. We apply phylogenetic bootstrapping, in combination with ordinary bootstrap resampling over cases, to calculate non-parametric confidence intervals on our measure of informational value.

For each bootstrap iteration, we first perform a resampling (with replacement) from the 29,903 sites of the SaRS-CoV-2 genome, which may repeat or omit sites from the original sequence, and construct a phylogenetic tree with evolutionary distances based on maximum likelihood. We consider two sequences to constitute a plausible transmission link if their distance (*i.e.*, expected number of substitutions) is fewer than two base pairs. The expected number of substitutions is the estimated substitution rate (observed mismatches divided by the number of non-missing sites) multiplied by the total number of sites. After obtaining evolutionary distances, we then resample cases (with replacement) for each bootstrap iteration to address sampling uncertainty.

Finally, for each iteration, we group all sequenced pairs of cases that are identified by a disease surveillance strategy and calculate that strategy’s informational value, equal to the fraction of proposed sequenced case pairs that are also plausible transmission links according to the bootstrapped tree. We repeat this process for 100 trials using parallelization, given one bootstrap alone requires nearly twenty-four hours of computation. From the bootstrap trials, we obtain a distribution of estimates for the informational value of each disease surveillance strategy, allowing us to calculate empirical confidence bounds and conduct statistical tests to determine whether a given strategy is significantly better than another. We report p-values when comparing two disease surveillance strategies, defined as the proportion of the 100 trials in which their ranking reverses, compared to their ranking based on average informational value.