SYMBOLIC TRANSFER ENTROPY REVEALS THE AGE STRUCTURE OF PANDEMIC INFLUENZA TRANSMISSION FROM HIGH-VOLUME INFLUENZA-LIKE ILLNESS DATA

STEPHEN M KISSLER[1,2], CÉCILE VIBOUD[3], BRYAN T GRENFELL[4], JULIA R GOG[1]

[1] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge, United Kingdom*

[2] *Department of Immunoloy and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America*

[3] *Fogarty International Center, National Institutes of Health, Bethesda, Maryland, United States of America*

[4] *Department of Ecology and Evolutionary Biology, University of Princeton, Princeton, New Jersey, United States of America*

ABSTRACT. Existing methods to infer the relative roles of age groups in epidemic transmission can normally only accommodate a few age classes, and/or require data that are highly specific for the disease being studied. Here, symbolic transfer entropy (STE), a measure developed to identify asymmetric transfer of information between stochastic processes, is presented as a way to determine which age groups drive an epidemic. STE provides a ranking of which age groups dominate transmission, rather than a reconstruction of the explicit between-age-group transmission matrix. Using simulations, we establish that STE can identify which age groups dominate transmission, even when there are differences in reporting rates between age groups and even if the data is noisy. Then, the pairwise STE is calculated between time series of influenza-like illness for 12 age groups in 884 US cities during the autumn of 2009. Elevated STE from 5-19 year-olds indicates that school-aged children were the most important transmitters of infection during the autumn wave of the 2009 pandemic in the US. The results may be partially confounded by higher rates of physician-seeking behaviour in children compared to adults, but it is unlikely that differences in reporting rates can explain the observed differences in STE.

KEYWORDS: Symbolic transfer entropy; pandemic influenza; age structure; electronic medical records; influenza-like illness

1. INTRODUCTION

Age is a key predictor of a person's rate of both acquiring [1, 2, 3, 4, 5, 6] and transmitting [7, 8, 9] influenza. Children tend to contribute more to influenza transmission than adults do [4, 7, 8], but the precise epidemiological roles of different age groups can shift from season to season [10] and may change markedly in pandemic years [11]. From a public health perspective, untangling the relative roles of different age groups could help guide targeted vaccination strategies [7, 12, 13, 14] and other age-related interventions, like the selective closure of schools [15, 16, 17]. However, data with sufficient resolution to identify detailed epidemiological relationships between age groups has so far been scarce, and even when such data exist, current methods are insufficient for reliably uncovering those relationships.

Electronic medical records (EMRs) help address the issue of data scarcity by providing high-volume influenza-like illness (ILI) incidence data with detailed age structure [18]. EMRs are routinely produced by physicians for insurance purposes during the majority of outpatient visits in the United States [18]. Since EMRs generally contain syndromic illness classifications, EMR-based estimates of influenza incidence are subject to noise from non-ILI respiratory infection. EMR-based disease incidence estimates are also subject to geographic and demographic variation in physician-seeking behaviour. Laboratory-confirmed influenza cases, as collected routinely by the Centers for Disease Control and Prevention (CDC) [19], provide more specific estimates of influenza incidence, but at substantially lower volume. Influenza incidence estimates from online search platforms and social media websites like Google [20] and Twitter [21] can provide massive amounts of data, but these sources' reliability has been called into question, and they lack detailed age information [22]. Dedicated online platforms such as FluNearYou in the US and FluSurvey in the UK, which gather reports of ILI symptoms from community volunteers [23, 24], hold some promise for supplementing traditional ILI data streams [25, 26, 27], but represent a relatively small convenience sample of the population. So, while other data sources exist, EMRs offer a relatively promising and so-far underutilised source of fine-scale data on influenza incidence in the United States [18, 22].

Previous attempts to infer the relative importance of different age groups for the transmission of influenza have sought to either reconstruct the explicit next-generation matrix (NGM) [3, 28, 29] or to infer the relative risk of infection between age groups [4]. The NGM-based methods have only been applied to scenarios with at most two age groups (children and adults), in part because they require strong assumptions about the structure of the next-generation matrix which become increasingly unrealistic as the number of age classes grows. The relative risk method [4] has been used to rank the importance of five age groups for the transmission of influenza, but requires data with high specificity for influenza, effectively precluding ILI datastreams and the use of EMRs in particular.

Symbolic transfer entropy (STE) [30] offers a way to infer the relative transmissive importance of possibly many age groups from ILI data. STE is an extension of transfer entropy (TE) [31], which measures the amount of information the past states of one stochastic process provide about the transition probabilities of another. Intuitively, the TE is a measure of the amount of information "transferred" from one stochastic process to another. To compute the STE, a time series is symbolised using a scheme that encodes its qualitative structure in a low-dimensional space, and then the TE is calculated from the relative frequencies of these symbols. The symbolisation scheme makes the STE robust to minor point-wise noise and to systematic shifts in amplitude, which in the context of EMR ILI data might arise from the presence of non-influenza ILI cases and from differences in reporting rate between age groups. These benefits come with the trade-off of requiring relatively large amounts of data compared to existing methods for inferring the age structure of disease transmission. STE has been used to study epileptogenic neural signals and the dissemination of information through social networks [30, 32], but to our knowledge has not been systematically evaluated as a means of providing insight into infectious disease transmission. TE and STE are similar to other model-free methods that measure shared information and so-called 'causal' relationships between stochastic processes, including mutual information [31], Granger causality [33], and convergent cross mapping [34]. Permutation entropy, a related measure, has recently been used to quantify the predictability of infectious disease outbreaks [35].

Here, we use influenza-like outbreak simulations to demonstrate that STE reliably identifies the age groups that drive influenza transmission. Then, we utilise an EMR-based dataset capturing ILI incidence from 884 ZIP (postal) codes and 12 age classes across the United States to rank the relative importance of the various age groups in the transmission of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in that country. We conclude that school-aged children (5–19 year-olds) were disproportionately responsible for transmitting influenza to infants through working-age adults in the autumn of 2009, in broad agreement with other findings. Our work demonstrates that STE could serve as an important tool for the detailed epidemiological analysis of age structure, especially as EMR data become more prevalent.

2. Materials and Methods

2.1. **Data.** The data come from a convenience sample of CMS-1500 electronic medical claims forms submitted by primary care physicians across the US and maintained by SDI health (now

IQVIA). Each claim is associated with a single outpatient visit, and includes one or more ICD-9 codes [36] listed by the physician that describe the patient's illness. The overall sample is thought to capture over 50% of all outpatient visits in the US in 2009 [18]. The records are binned weekly and aggregated geographically by the first three digits of the ZIP (postal) code of the practice from which they are submitted [37]. These three-digit ZIP codes will be referred to simply as 'ZIPs' (not to be confused with the finer five- or ten-digit ZIP codes, also assigned to many mailing addresses in the US [36]). Time series of weekly influenza-like illness (ILI) incidence are created by extracting claims with a direct mention of influenza, or fever combined with a respiratory symptom, or febrile viral illness (ICD-9 487-488 OR [780.6 and (462 or 786.2)] OR 079.99), following Viboud *et al.* (2014) [18]. For each ZIP, the number of ILI cases in each week is divided by the total number of patients who visited a physician in that ZIP during that week, yielding an 'ILI ratio' time series. There are 884 ILI ratio time series, one for each ZIP in the lower 48 US states, each spanning 52 weeks from the week commencing 4 Jan 2009 through the week commencing 27 Dec 2009. The correspondence between the SDI-ILI dataset and reference influenza surveillance data from the US Centers for Disease Control and Prevention (CDC) is described in depth by Viboud *et al.* (2014) [18].

2.2. **Symbolic transfer entropy.**

If I and J are discrete-state and discrete-time random processes such that i_t and j_t are the states of processes I and J at time t, then the transfer entropy (TE) from process J to process I is defined as

$$T_{J\to I} = \sum_{\Omega_I,\Omega_J} p(i_{t+1}, i_t^{(k)}, j_t^{(l)}) \log\Big(\frac{p(i_{t+1}|i_t^{(k)}, j_t^{(l)})}{p(i_{t+1}|i_t^{(k)})}\Big) \tag{1}$$

where $i_t^{(k)}$ is shorthand notation for the k-step history of process i, $(i_t, \ldots, i_{t-k+1})$, and similarly $j_t^{(l)} = (j_t, \ldots, j_{t-l+1})$. The logarithm has base 2, so that the transfer entropy is measured in bits. The sum is over all possible combinations of states $(i_{t+1}, i_t^{(k)}, j_t^{(l)})$, where $i_{t+1}, i_t^{(k)} \in \Omega_I$ and $j_t^{(l)} \in \Omega_J$, and Ω_I and Ω_j are the state spaces for processes I and J. Eq. 1 is a Kullback-Leibler divergence that measures how much process I deviates from the generalised Markov property $p(i_{t+1}|i_t, \ldots, i_1) = p(i_{t+1}|i_t^{(k)})$, given the last l states of process J. In practice, the histories are often fixed at length 1 ($k = l = 1$) and the probabilities are estimated from simple counts of the observed data [31].

The TE is limited in that it is only defined for stochastic processes with a discrete state space. Staniek and Lehnertz (2008) [30] introduce symbolic transfer entropy (STE) as a way to calculate information transfer between time series processes that have continuous- or near-continuous state spaces. Motivated by the insight that the relative amplitudes of subsequent observations from these sorts of processes may provide enough information to reveal interactions between them, they propose symbolising the time series based on ordered m-tuples of observations (Fig. S1). This reduces the (near-)continuous state space of the original stochastic process to a discrete set of $m!$ symbols. In practice, m is often chosen to be 2 or 3, giving a state space of 2 or 6 symbols, respectively. For $m = 3$, we also tested the effect of collapsing the two concave-up and the two concave-down symbols into a single symbol each, resulting in a smaller state space (four *vs.* six symbols) while capturing a similar level of qualitative detail. Details on the symbolisation of time series and the empirical calculation of the STE are provided in the Supplemental Information.

2.3. **SIR epidemic simulation model.** For simulations with just two age classes, we use a stochastic SIR model implemented using the Gillespie algorithm [38]. For all simulations, the basic reproduction number R_0 is set at 1.5, consistent with estimates of the basic reproduction number of 2009 A/H1N1 pandemic influenza [39, 40]. We consider a population size of $N = 1,000$ split evenly between classes 1 and 2, so that $N_1 = N_2 = 500$ (age groups with different population sizes are also considered in the Supplemental Information). The expected time to recovery $1/\gamma$ is assumed constant for all age groups and is set at 7 days, which is consistent with estimates of the infectious period for 2009 pandemic influenza [40]. Table S1 gives the rates at which individuals of each class stochastically progress from susceptible to infected to recovered. Infections are binned into week-long intervals, and Poisson noise is added to simulate non-influenza influenza-like illness. Fig. S7 depicts five incidence time series produced using the model. Full details on the model and simulation procedure are given in the Supplemental Information.

2.4. **Poisson epidemic simulation model.** For more than two age classes, the full stochastic SIR model becomes too computationally demanding for repeated simulations to be practical. So, we also define an outbreak simulation model based on a self-exciting Poisson process, similar to [41]. We choose the time units t to match the mean generation interval of the infection, which we set at 3.5 days [42]. To generate epidemics, we use a stepwise-constant effective reproduction number R_t, such that $R_t = 1.5$ for the first four weeks (eight generations) of the outbreak and

$R_t = 0.8$ thereafter. Infections are binned from the half-week generations into week-long intervals, and additional Poisson noise is added to each bin to simulate non-influenza influenza-like illness. For simulations with two age classes, the Poisson model yields epidemics of similar length and magnitude as the two-age-class SIR model (compare Figs S7 and S8), and yields comparable STE inferences (see Fig. 1), which suggests that the Poisson model is an acceptable approximation to the stochastic SIR model. Full details on the implementation of the Poisson model are given in the Supplemental Information.

2.5. **Reporting rates.** Only a fraction of influenza cases are represented in the SDI-ILI dataset, since many people do not seek medical care for their symptoms. The tendency to seek medical care given infection with an ILI can vary by age group [43]. To factor this into the outbreak simulations, we introduce a reporting rate vector $\boldsymbol{c}$ in which element c_i gives the expected proportion of individuals in age class i who seek medical care when infected with an ILI. It is then possible to simulate a 'reported' disease incidence time series:

$$Y_{i,t}^{obs} \sim Binomial(Y_{i,t}, c_i) \tag{2}$$

where $Y_{i,t}$ is the simulated number of infected individuals in age class i at time t (under either model) and $Y_{i,t}^{obs}$ is the simulated reported number of infections in age class i at time t.

3. Results

3.1. **STE reveals transmission asymmetries between two coupled age groups.** We first calculate the STE between two age groups as the within- and between-group reproduction ratios vary. We consider between-group transmission that ranges from (a) fully decoupled to fully symmetric, and (b) fully symmetric to strongly driven by Group 1. The between-group infectiousness is specified using a "relative reproduction matrix" $\boldsymbol{r}$, which is a scaled version of the next-generation matrix [28], such that $\boldsymbol{r}_{i,j}/\boldsymbol{r}_{k,j}$ gives the proportional difference in group j's infectiousness for group i *vs.* group j. For example, if $\boldsymbol{r}_{i,j}/\boldsymbol{r}_{k,j} = 2$, then a member of group j is expected to infect twice as many members of group i than of group k. Scenario (a) is encapsulated by the relative reproduction matrix

$$\boldsymbol{r_a} = \begin{bmatrix} 1 & z_a \\ z_a & 1 \end{bmatrix} \tag{3}$$

where $z_a \in [0, 1]$. Scenario (b) is encapsulated by the relative reproduction matrix

$$\boldsymbol{r_b} = \begin{bmatrix} 1 + 3z_b & 1 \\ 1 + z_b & 1 \end{bmatrix} \tag{4}$$

where $z_b \in [0, 1]$.

Fig. 1 depicts the change in STE under these two transmission scenarios, calculated from epidemics simulated using the stochastic SIR model (Fig. 1 A–B) and the Poisson model (Fig. 1 C–D). Each pane in Fig. 1 is produced using 100 ensembles of 800 simulated epidemics for each value of z_a and z_b between 0 and 1 in steps of size 0.1. For each ensemble, the 800 simulated incidence time series are symbolised using symbols of length $m = 3$, and then the between-group transfer entropies are estimated using the relative symbol frequencies (see Fig. S3), producing 100 STE estimates for each value of z_a and z_b. The solid blue (black) lines in Fig. 1 depict the mean Group 1→2 (Group 2→1) STE for each value of z_a and z_b across the 100 ensembles. The shaded blue (black) bands depict the range of the middle 95 Group 1→2 (Group 2→1) STE estimates for each value of z_a and z_b across the 100 ensembles, analogous to a 95% confidence interval. Under both the stochastic SIR and the Poisson models, the between-group STE increases steadily as the transmissive coupling ranges from none to symmetric (Fig. 1 A, C). Once Group 1 begins to dominate transmission, the Group 1→2 STE increases and the Group 2→1 STE decreases (Fig. 1 B, D), accurately capturing the transmissive relationship between the age groups.

When Group 1 drives transmission, the Poisson model yields a smaller difference in the STE between the two age groups than the stochastic SIR model does (Fig. 1 B, D). Visual inspection suggests that the simulated time series produced using the stochastic SIR model tend to feature more stochastic fluctuations than the time series produced using the Poisson model (Figs S7 and S8). Since STE is effectively a measure of how these stochastic fluctuations transmit from one age group to another, this may explain why the differences in STE calculated using the Poisson model are relatively less pronounced. Overall, the qualitative similarity between the STE estimates from the two transmission models suggests that the Poisson model is an acceptable approximation to the stochastic SIR model, and that simulations from the Poisson model tend to produce more conservative estimates of the difference in STE between age groups than the stochastic SIR model.

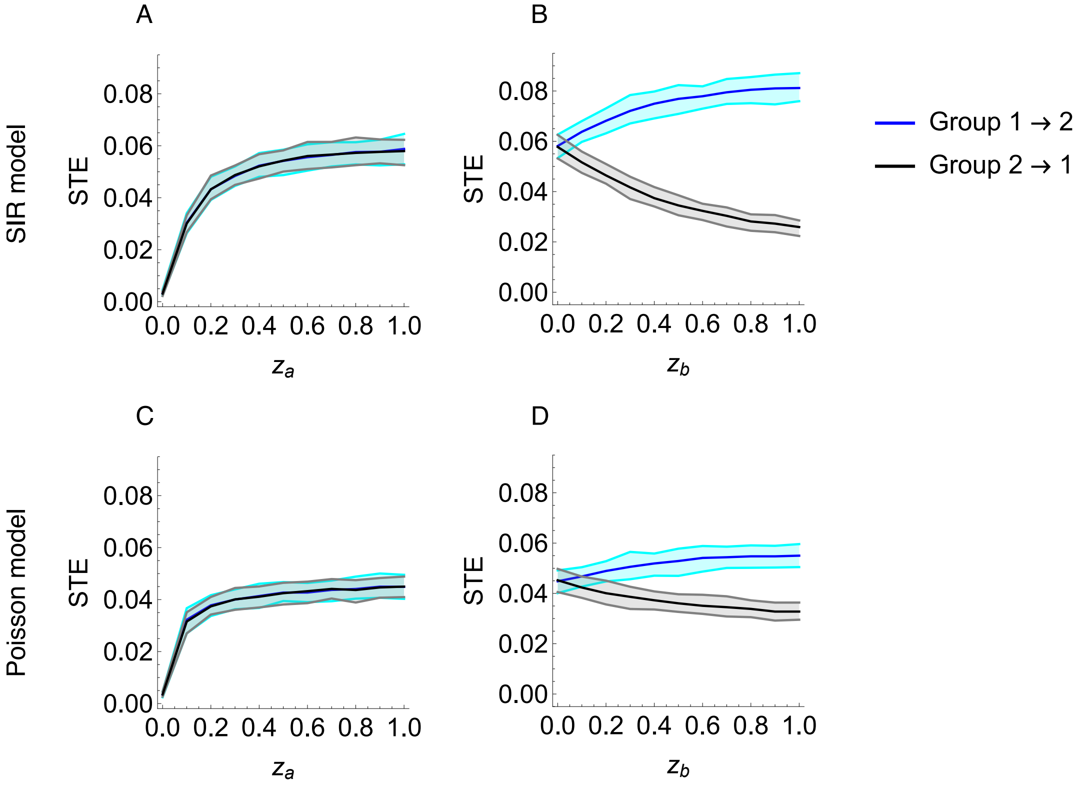

FIGURE 1. Mean (95% CI) Group 1→2 (blue) and Group 2→1 (black) STE values as the coupling between the two groups ranges from none to fully symmetric (A and C), and from fully symmetric to strongly driven by Group 1 (B and D). The curves are produced by simulating 100 ensembles of 800 epidemics each from the stochastic SIR model (A and B) or the Poisson model (C and D) for each value of z_a and z_b between 0 and 1 in steps of 0.1, and then calculating the between-group STE for each ensemble. The relative reproduction matrices that capture these two coupling scenarios are given in Eqs 3 and 4.

3.2. **STE reveals transmission asymmetries despite incomplete reporting.** Next, we evaluate how incomplete reporting influences the detection of asymmetries in transmission strength. Fig. 2 depicts the mean estimated STE across 100 ensembles of 800 epidemics each for reporting rates c_i between 0.1 and 1 in steps of 0.1, with equal reporting rates across all age groups. The epidemic simulations are produced using the Poisson model with relative reproduction matrix

$$\boldsymbol{r} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 4 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{5}$$

which could represent ‘children’ (Group 2) having strong within-group transmission ($r_{2,2} = 4$) and intermediate transmission to ‘infants’ (Group 1) and ‘adults’ (Group 3) ($r_{1,2} = r_{3,2} = 2$). Even for reporting rates as low as 0.1, the STE values from Group 2 are higher than those from any other group. As the reporting rates increase, the differences become more pronounced, accurately capturing the transmissive dominance of Group 2 over the other groups. The estimated STE increases with reporting rate for all age groups, but more quickly for Group 2 than for the other age groups. According to Biggerstaff *et al.* (2012) [43], true reporting rates for ILI in the US during the 2009 pandemic were between 0.4 and 0.6, for which the transmissive dominance of Group 2 is clear.

3.3. **STE reveals transmission asymmetries between twelve coupled age groups.** To test the ability of STE to identify transmission asymmetries from data on the scale of the SDI-ILI dataset, we use the Poisson model to simulate 100 ensembles of 800 epidemics each with 12 age groups. We consider the scenarios (a) with the 12×12 relative reproduction matrix Eq. S48, representing high transmission from Groups 3–5 to Groups 3–5 ($r_{i,j} = 4$ for $i, j \in \{3, 4, 5\}$), intermediate transmission from groups 3–5 to groups 1–2 and 6–9 ($r_{i,j} = 2$ for $i \in \{1, 2, 6, 7, 8, 9\}$ and $j \in \{3, 4, 5\}$), baseline transmission ($r_{i,j} = 1$) between all other groups, and uniform 50% reporting rate across all groups, and (b) with uniform transmission strength across all age groups (i.e. a 12×12 relative reproduction matrix with ‘1’ for all entries), 60% reporting rate for groups 1–5, and 40% reporting rate for groups 6–12, following the estimates of Biggerstaff *et al.* (2012) [43] for the ILI reporting rates in the United States during the 2009 influenza pandemic for children and adults, respectively.

Fig. 3 depicts the mean pairwise STE estimates between the 12 age groups under both scenarios. The square in row i and column j represents the STE from Group j to Group i. Darker squares correspond to higher STE. For the asymmetric transmission/uniform reporting rate scenario (scenario (a), Fig. 3A), the STE clearly captures the transmissive dominance of Groups 3,

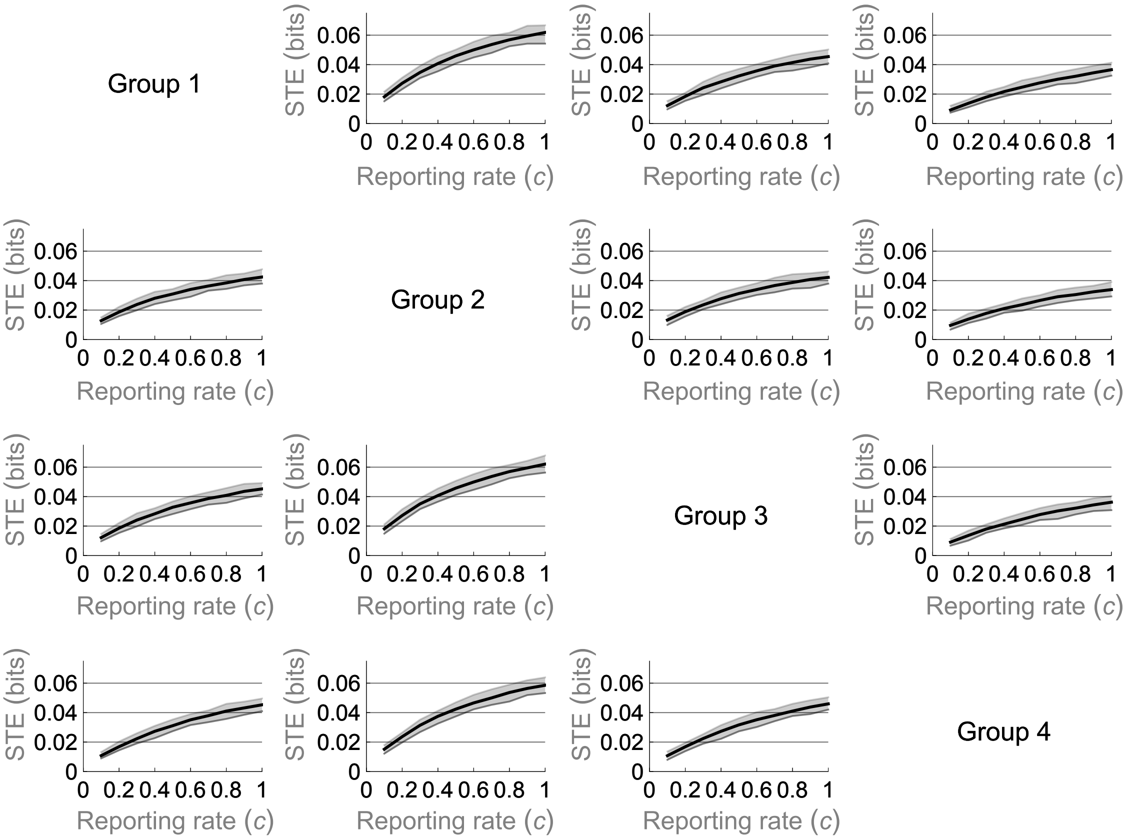

FIGURE 2. Mean pairwise STE values (solid lines) with 95% confidence intervals (shaded bands) for epidemics strongly driven by Group 2 under a range of reporting rates c. The curves are produced by simulating 100 ensembles of 800 epidemics each from the Poisson model for each value of c between 0.1 and 1 in steps of 0.1, and then calculating the between-group STE for each ensemble. The reporting rate c_i (see Eq. 2) is varied uniformly across all age groups i. The relative reproduction matrix that specifies within- and between-group transmission rates is given by Eq. 5. The plot in row i and column j depicts the STE from group j to group i.

4, and 5. The pairwise STE does not simply reproduce the structure of the relative reproduction matrix, as evidenced by the variability in mean pairwise STE for age groups other than Groups 3–5. This is because the STE captures a ‘knock-on’ effect for which information transferred from a strongly-driving age group can propagate through other age groups. For the uniform transmission/variable reporting rate scenario (scenario (b), Fig. 3B), it is evident that elevated reporting rates can also lead to elevated STE, both to and from the groups with elevated reporting rate (Groups 1–5). Overall, the variability in STE due to differences in reporting rate appears to be smaller than the variability in STE due to differences in transmission strength. Further discussion on the effect of reporting rates on STE may be found in the Supplemental Information.

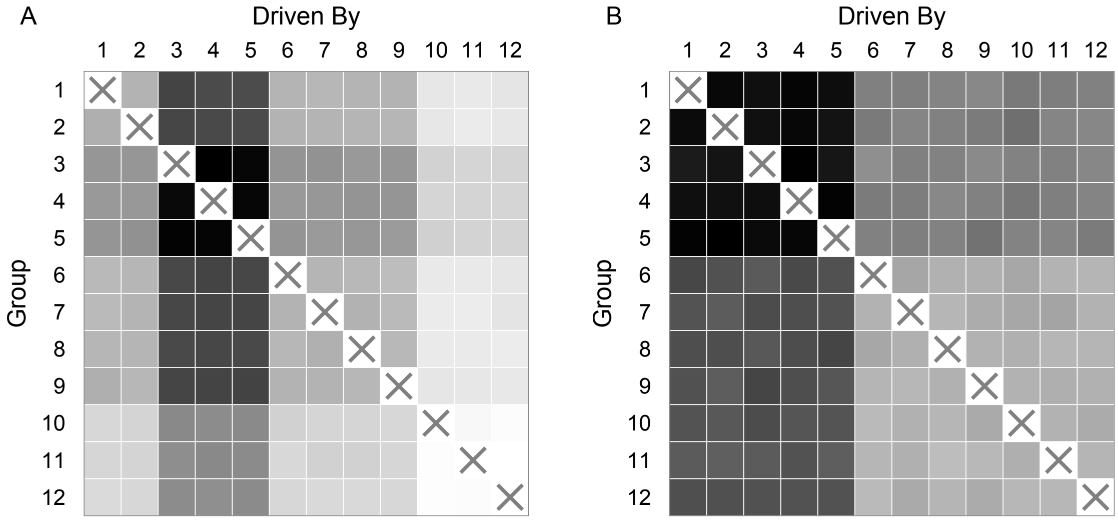

FIGURE 3. Mean pairwise STE values between 12 groups for epidemics strongly driven by Groups 3, 4, and 5 and uniform 50% reporting rate across all age groups (A), and for epidemics driven equally by all age groups, 60% reporting for Groups 1–5, and 40% reporting for Groups 6–12 (B). A box in row i and column j corresponds to the STE from group j to group i, where darker shades corresponds to higher STE. To generate the STE values, 100 ensembles of 800 epidemics were simulated from the Poisson model using relative rate matrix Eq. S48 for (A) or a relative rate matrix with all entries equal to 1 for (B). Each ensemble generates 144 pairwise STE values, so that each box represents the mean value across the 100 ensembles. The raw values are listed in Eqs S49 and S50.

3.4. **School-aged children contributed disproportionately to transmission during the autumn 2009 A/H1N1pdm influenza outbreak in the US.** To estimate the pairwise STE between the 12 age groups represented in the SDI-ILI dataset during the 2009 A/H1N1pdm influenza pandemic, we extract data from the 25 weeks between 12 July 2009 and 27 December 2009 and symbolise the ILI time series for each age group in each ZIP using a symbol length of $m = 3$. The pairwise STE values between all age groups are depicted in Fig. 4. The STE is highest in the columns representing 5–19 year-olds. This provides evidence that there was systematically elevated transmission from school-aged children to infants through adults. The adult-adult STE is also moderately elevated, suggesting that adults may have played a relatively important role in transmitting the outbreak amongst themselves, though this could also be explained by elevated transmission from children alone. Compare, for example, to the left-hand plot in Fig 3: in that simulation, only transmission from children is elevated, but it causes a moderate elevation in the STE from adults and infants to the other age groups due to the knock-on effect.

As a control, we also calculated the pairwise STE between all age groups during 25 post-pandemic weeks, from 10 January 2010 through 27 June 2010. For these months, there is no apparent age structure in transmission (see Supplemental Information). We also calculated the pairwise STE between age groups for six previous influenza seasons (see Supplemental Information). For the 2009 pandemic, there is is a higher maximum pairwise STE and greater variation in the pairwise STEs than for any previous season. This could reflect differences in baseline ILI, which was likely lower during the autumn 2009 pandemic wave than during the seasonal outbreaks, due to the pandemic's earlier timing. A lower baseline ILI might have made pairwise differences in STE easier to detect in 2009. However, the relatively higher and more heterogeneous STE values in 2009 are also consistent with the hypothesis that school-aged children played a disproportionately large role in the spread of the 2009 pandemic, as has been described elsewhere [4].

It is unlikely that differences reporting rates alone can account for the elevated STE from 5–19 year-olds to the other age groups. The mean pairwise STE values computed from simulations with uniform transmission rates and unequal reporting rates in Section 3.3 range from .0057 to 0.0084 (see Eq. S50), while the pairwise STE values computed from the SDI-ILI data range from 0.0056 to 0.084 (see Eq. S51), an order of magnitude larger. The mean pairwise STE values computed from simulations with asymmetric transmission and uniform reporting rates Section 3.3 range from 0.0047 to 0.014 (see Eq. S49), closer to the range observed from the SDI-ILI data but still somewhat smaller. This points towards a possible combined effect of strong transmissive driving from children plus elevated reporting in children. In addition, re-calculating the pairwise STE using probabilistic reconstructions of the pre-reporting SDI-ILI incidence time series (see Supplemental Information) indicate that the observed transmissive dominance of 5–19 year-olds persists even after adjusting for potential differences in reporting rate between children and adults. Furthermore, Biggerstaff *et al.* (2012) [43] report that 0-4 year-olds had the highest reporting rates for ILI in the United States in 2009, yet the STE from 0-4 year-olds is relatively low compared to the other age groups. If reporting rates alone could explain the observed differences in STE, the STE from infants should be at least as high as the STE from school-aged children.

It is also unlikely that the unequal partitions of the age groups can explain the observed patterns in the pairwise STE. The age groups under 20 years are partitioned such that they span fewer years, and thus contain fewer individuals, than the age groups above 20 years. Direct calculations and simulations (see Supplemental Information) indicate that, all else being equal, the out-going STE for a given group tends to increase as the group's population size increases relative to the sizes of the other groups. If differences in the groups' population sizes were driving the observed pairwise STE values, we would expect the age groups over 20 years to appear to dominate transmission – which is the opposite of what we observe here.

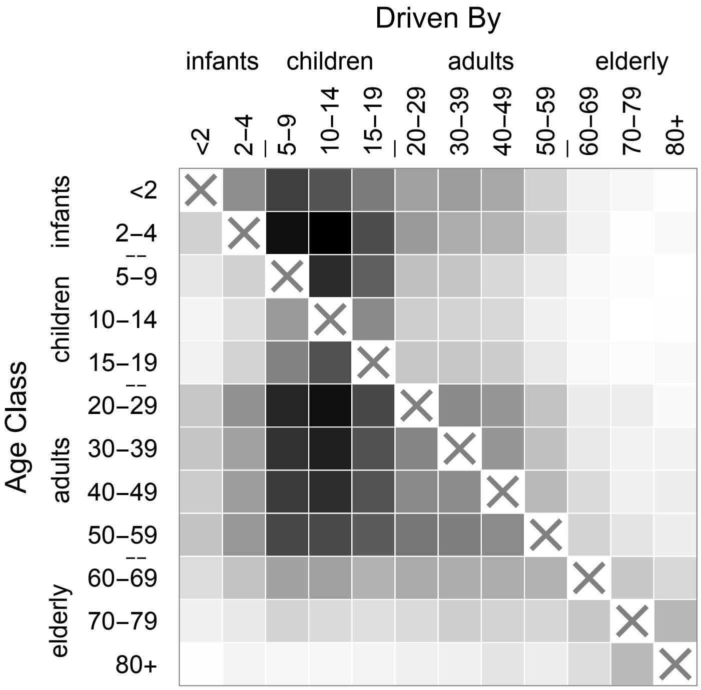

FIGURE 4. Mean pairwise STE values between the 12 groups represented in the SDI-ILI dataset during the autumn 2009 A/H1N1pdm pandemic influenza outbreak. A box in row i and column j corresponds to the STE from group j to group i, where darker shades corresponds to higher STE. The raw values are listed in Eq. S51.

4. DISCUSSION

Here, we propose STE as a means of ranking which age groups contribute most to the transmission of infectious disease outbreaks. STE is chosen for its robustness to point-wise noise and overall amplitude shifts in time series, which especially affect the ILI data stream due to non-influenza respiratory illness and incomplete reporting. Simulation studies indicate that STE can correctly rank transmissive asymmetries between age groups. However, STE is also positively associated with reporting rates, which can partially confound estimates of asymmetric transmission.

STE estimates from ILI time series data from July-December 2009 in the United States suggest that 5–19 year-olds were primarily responsible for driving transmission of the autumn wave of the A/H1N1pdm pandemic influenza outbreak. It is unlikely that this result can be explained by differences in reporting rates alone.

The identification of school-aged children as the primary drivers of transmission of the 2009 influenza pandemic in the United States agrees with most other studies on age-specific transmission of both seasonal and pandemic influenza [4, 7, 8, 9]. Elevated transmission from school-aged children is likely due in part to the relatively high number of daily interpersonal contacts made by members of these age groups. Mossong *et al.* (2008) [8] for example estimate that 10–19 year-olds have more contacts per day than any other age group, and conclude from a modelling study based on empirical contact data that 5–19 year-olds are likely to both suffer the highest burden of disease and to drive the early-stage transmission of an outbreak transmitted by droplets through close contacts, like influenza. This underscores the importance of monitoring children during pandemic influenza outbreaks, and potentially prioritizing school-aged children for vaccination.

TE is closely linked to mutual information [31] and Granger causality [33]. Unlike TE, mutual information is symmetric; that is, it measures the probabilistic dependence between two processes, but cannot determine the direction of information transfer between them, if there is any [31]. Measuring the delayed mutual information between two processes is one way to introduce asymmetry. This takes a step toward inferring whether one process influences another, by measuring shared information between the present state of one process and the past states of another [31]. While the lagged mutual information describes how one process' history predicts the static probabilities of another, the TE measures how one process' history influences the transition probabilities of another. Because of this, the TE is less likely to be confounded by a shared input signal, and is a better measure of stochastic 'driving' [31]. Section 2 of Kaiser and Schreiber (2002) [44] provides a detailed description of the differences between TE and mutual information. Granger causality, on the other hand, is a special case of TE that arises when the stochastic processes are jointly Gaussian-distributed [45]. The TE is thus better suited than Granger causality for making inferences on more general, possibly nonlinear, processes, though this comes at the expense of requiring more data and having no clear way to test statistical significance [45].

Convergent cross mapping (CCM) [34] was developed to solve a similar problem as TE, but is based on somewhat different underlying theory. CCM was developed to detect so-called 'causal' relationships in partially stochastic systems with underlying deterministic structure. CCM relies on Takens' theorem [46] to reconstruct candidate manifolds of the underlying dynamical system using lagged observations from two time series. 'Causality' is inferred if nearby points on one reconstructed manifold consistently map to nearby points on the other reconstructed manifold. CCM has been used to provide evidence that temperature and absolute humidity fluctuations drive the timing of global seasonal influenza outbreaks [47], though some controversy surrounds these findings [48, 49]. Nevertheless, it would be interesting to see whether CCM can reveal asymmetric epidemiological interactions between age groups, and to compare its findings with those identified using TE. Lungarella *et al.* (2007) [50] provide more detail on the relationships between various methods that infer asymmetric relationships from time series data. (As an aside, we prefer to avoid the term 'causality' with respect to these methods, despite its frequent use in the literature. Regardless of the vocabulary used, they have successfully detected meaningful relationships between real-world coupled dynamic processes [30, 32, 34, 51, 52, 53]).

Despite the apparent well-suitedness of STE for making inferences from ILI data, its epidemiological relevance currently remains limited. The calculation of STE requires no prior epidemiological information whatsoever, which makes its success somewhat surprising. The next-generation matrix [28] is the key object for characterising age-structured, or more generally population-structured, disease transmission dynamics, and yet there is no obvious direct link between STE estimates and the NGM. It is possible that further simulation studies could help identify such a link; even though the STE values seem to bear little mechanistic meaning apart from the relative ordering of age groups that they yield, it is possible that regressing the inferred STE values on an underlying known NGM could connect the pairwise STE matrix with the NGM under certain conditions. However, it appears unlikely that a simple link exists, especially since STE can say nothing about transmission within a single age group, which is necessary for filling in the diagonal entries of the NGM. STE and related methods such as CCM that do not explicitly incorporate mechanistic descriptions of the underlying physical system are unlikely to be able to reveal more than an approximate hierarchy of driving processes. Nevertheless, such a hierarchy can contain valuable information, especially if developing and fitting a mechanistic model is too demanding to be practicable. Certain extensions to STE could also enhance its relevance for epidemiological inference. Local transfer entropy [54] and state-dependent transfer entropy [55], like the contextual STE (see Supplemental Information), are intended to make the TE more flexible and general, by considering how information transfer may change under varying conditions or ‘meta-states’. These extensions may yield better insight into epidemic processes, which are inherently nonlinear and context-dependent, than the more traditional measurements of transfer entropy can provide.

Perhaps the most important challenge confronting the TE and related measurements is deciding how to measure statistical power and significance. STE calculations rely on a middle level of stochasticity in the underlying stochastic processes; for a deterministic system, the STE will always be exactly zero, while for a stochastic system with too much within-sequence noise, the small-scale variation in amplitudes will likely mask important patterns from which the transfer of information might be inferred. The acceptable range of stochasticity has not been clearly defined. Similarly, it is unclear how best to measure when a difference in STE should be called statistically significant. Though this is recognised as an open and difficult problem [45, 48], it may be possible to make some progress by assuming that the underlying process follows certain epidemiological, or otherwise well-specified, dynamics.

5. Disclaimer

This paper does not necessarily represent the views of the US government or the NIH.

6. Funding

SK was supported by a Gates Cambridge scholarship.

References

[1] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds, “Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases,” *PLoS Medicine*, vol. 5, p. e74, mar 2008.

[2] C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. M. E. Guevara, F. Checchi, E. Garcia, S. Hugonnet, and C. Roth, “Pandemic potential of a strain of influenza A (H1N1): early findings.,” *Science (New York, N.Y.)*, vol. 324, no. 5934, pp. 1557–61, 2009.

[3] H. Nishiura, C. Castillo-Chavez, M. Safan, and G. Chowell, “Transmission potential of the new influenze A(H1N1) virus and its age-specificity in Japan,” *Eurosurveillance*, vol. 14, no. 22, pp. 1–5, 2009.

[4] C. J. Worby, S. S. Chaves, J. Wallinga, M. Lipsitch, L. Finelli, and E. Goldstein, “On the relative role of different age groups in influenza epidemics,” *Epidemics*, vol. 13, pp. 10–16, 2015.

[5] Y. Yang, J. D. Sugimoto, M. E. Halloran, N. E. Basta, D. L. Chao, L. Matrajt, G. Potter, E. Kenah, and I. M. Longini, “The transmissibility and control of pandemic influenza A(H1N1) virus,” *Science*, vol. 326, no. 5953, pp. 729–733, 2009.

[6] J. S. Brownstein, K. P. Kleinman, and K. D. Mandl, “Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system,” *American Journal of Epidemiology*, vol. 162, no. 7, pp. 686–693, 2005.

[7] J. Wallinga, P. Teunis, and M. Kretzschmar, “Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents,” *American Journal of Epidemiology*, vol. 164, no. 10, pp. 936–944, 2006.

[8] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds, “Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases,” *PLoS Medicine*, vol. 5, p. e74, mar 2008.

[9] T. Smieszek, M. Balmer, J. Hattendorf, K. W. Axhausen, J. Zinsstag, and R. W. Scholz, “Reconstructing the 2003/2004 H3N2 influenza epidemic in Switzerland with a spatially explicit, individual-based model,” *BMC Infectious Diseases*, vol. 11, no. 1, p. 115, 2011.

[10] T. Bedford, S. Riley, I. G. Barr, S. Broor, M. Chadha, N. J. Cox, R. S. Daniels, C. P. Gunasekaran, A. C. Hurt, A. Kelso, A. Klimov, N. S. Lewis, X. Li, J. W. McCauley, T. Odagiri, V. Potdar, A. Rambaut, Y. Shu, E. Skepner, D. J. Smith, M. a. Suchard, M. Tashiro, D. Wang, X. Xu, P. Lemey, and C. a. Russell, “Global circulation patterns of seasonal influenza viruses vary with antigenic drift,” *Nature*, vol. 523, no. 7559, pp. 217–20, 2015.

[11] M. A. Miller, C. Viboud, M. Balinska, and L. Simonsen, “The Signature Features of Influenza Pandemics Implications for Policy,” *New England Journal of Medicine*, vol. 360, no. 25, pp. 2595–2598, 2009.

[12] I. M. Longini and M. E. Halloran, “Strategy for Distribution of Influenza Vaccine to High-Risk Groups and Children,” *American Journal of Epidemiology*, vol. 161, no. 4, pp. 303–306, 2005.

[13] S. D. Mylius, T. J. Hagenaars, A. K. Lugnér, and J. Wallinga, “Optimal allocation of pandemic influenza vaccine depends on age, risk and timing,” *Vaccine*, vol. 26, pp. 3742–3749, jul 2008.

[14] T. A. Reichert, N. Sugaya, D. S. Fedson, W. P. Glezen, L. Simonsen, and M. Tashiro, “The Japanese experience with vaccinating schoolchildren against influenza,” *N Engl J Med*, vol. 344, no. 12, pp. 889–896, 2001.

[15] V. Gemmetto, A. Barrat, and C. Cattuto, “Mitigation of infectious disease at school: Targeted class closure vs school closure,” *BMC Infectious Diseases*, vol. 14, no. 1, pp. 1–10, 2014.

[16] G. J. Milne, N. Halder, and J. K. Kelso, “The cost effectiveness of pandemic influenza interventions: a pandemic severity based analysis.,” *PloS one*, vol. 8, no. 4, p. e61504, 2013.

[17] S. Cauchemez, N. M. Ferguson, C. Wachtel, A. Tegnell, G. Saour, B. Duncan, and A. Nicoll, “Closure of schools during an influenza pandemic,” *The Lancet Infectious Diseases*, vol. 9, no. 8, pp. 473–481, 2009.

[18] C. Viboud, V. Charu, D. Olson, S. Ballesteros, J. Gog, F. Khan, B. Grenfell, and L. Simonsen, “Demonstrating the Use of High-Volume Electronic Medical Claims Data to Monitor Local and Regional Influenza Activity in the US,” *PLoS ONE*, vol. 9, p. e102429, jan 2014.

[19] Centers for Disease Control and Prevention, “Overview of Influenza Surveillance in the United States,” tech. rep., Centers for Disease Control and Prevention, 2016.

[20] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data.,” *Nature*, vol. 457, pp. 1012–4, feb 2009.

[21] V. Lampos, T. D. Bie, and N. Cristianini, “Flu Detector - Tracking Epidemics on Twitter,” in *Machine Learning and Knowledge Discovery in Databases* (J. Balcazar, F. Bonchi, M. Sebag, and A. Gionis, eds.), pp. 599–602, Berlin/Heidelberg: Springer-Verlag, 2010.

[22] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, “Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales,” *PLoS Computational Biology*, vol. 9, p. e1003256, jan 2013.

[23] HealthMap, “FluNearYou,” 2017.

[24] London School of Hygiene and Tropical Medicine, “FluSurvey,” 2017.

[25] A. J. Adler, K. T. Eames, S. Funk, and W. J. Edmunds, “Incidence and risk factors for influenza-like-illness in the UK: online surveillance using Flusurvey,” *BMC Infectious Diseases*, vol. 14, no. 1, p. 232, 2014.

[26] D. Perrotta, A. Bella, C. Rizzo, and D. Paolotti, “Participatory online surveillance as a supplementary tool to sentinel doctors for influenza-like illness surveillance in Italy,” *PLoS ONE*, vol. 12, no. 1, pp. 1–15, 2017.

[27] M. S. Smolinski, A. W. Crawley, K. Baltrusaitis, R. Chunara, J. M. Olsen, O. Wójcik, M. Santillana, A. Nguyen, and J. S. Brownstein, “Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons,” *American Journal of Public Health*, vol. 105, no. 10, pp. 2124–2130, 2015.

[28] O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts, “The construction of next-generation matrices for compartmental epidemic models,” *Journal of The Royal Society Interface*, vol. 7, no. 47, pp. 873–885, 2010.

[29] K. Glass, G. N. Mercer, H. Nishiura, E. S. McBryde, and N. G. Becker, “Estimating reproduction numbers for adults and children from case data,” *Journal of The Royal Society Interface*, vol. 8, pp. 1248–1259, sep 2011.

[30] M. Staniek and K. Lehnertz, “Symbolic Transfer Entropy,” *Phys. Rev. Lett.*, vol. 100, no. April, p. 158101, 2008.

[31] T. Schreiber, “Measuring Information Transfer,” *Physical Review Letters*, vol. 85, no. 2, p. 19, 2000.

[32] J. Borge-Holthoefer, N. Perra, B. Goncalves, S. Gonzalez-Bailon, A. Arenas, Y. Moreno, and A. Vespignani, “The dynamics of information-driven coordination phenomena: A transfer entropy analysis,” *Science Advances*, vol. 2, pp. e1501158–e1501158, apr 2016.

[33] C. W. J. Granger, “Investigating Causal Relations by Econometric Models and Cross-spectral Methods,” *Econometrica*, vol. 37, p. 424, aug 1969.

[34] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, “Detecting causality in complex ecosystems.,” *Science*, vol. 338, pp. 496–500, oct 2012.

[35] S. V. Scarpino and G. Petri, “On the predictability of infectious disease outbreaks,” *Nature Communications*, vol. 10, no. 1, 2019.

[36] I. M. Moriyama, R. M. Loy, and A. H. Robb-Smith, “History of the statistical classification of diseases and causes of death,” tech. rep., Centers for Disease Control and Prevention, 2011.

[37] U.S. Postal Service Office of Inspector General, “The Untold Story of the ZIP Code,” tech. rep., United State Postal Services, 2013.

[38] D. T. Gillespie, “Exact stochastic simulation of coupled chemical reactions,” *Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[39] M. A. Jhung, D. Swerdlow, S. J. Olsen, D. Jernigan, M. Biggerstaff, L. Kamimoto, K. Kniss, C. Reed, A. Fry, L. Brammer, J. Gindler, W. J. Gregg, J. Bresee, and L. Finelli, “Epidemiology of 2009 pandemic influenza a (H1N1) in the United States,” *Clinical Infectious Diseases*, vol. 52, no. SUPPL. 1, pp. 13–26, 2011.

[40] W. Yang, M. Lipsitch, and J. Shaman, “Inference of seasonal and pandemic influenza transmission dynamics,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 9, p. 201415012, 2015.

[41] L. Held, M. Hofmann, M. Höhle, and V. Schmid, “A two-component model for counts of infectious diseases,” *Biostatistics*, vol. 7, no. 3, pp. 422–437, 2006.

[42] P.-Y. Boëlle, S. Ansart, A. Cori, and A.-J. Valleron, “Transmission parameters of the A/H1N1 (2009) influenza virus pandemic: a review,” *Influenza and Other Respiratory Viruses*, vol. 5, pp. 306–316, sep 2011.

[43] M. Biggerstaff, M. Jhung, L. Kamimoto, L. Balluz, and L. Finelli, “Self-reported influenza-like illness and receipt of influenza antiviral drugs during the 2009 pandemic, United States, 2009-2010,” *American Journal of Public Health*, vol. 102, no. 10, pp. 2009–2010, 2012.

[44] A. Kaiser and T. Schreiber, “Information transfer in continuous processes,” *Physica D: Nonlinear Phenomena*, vol. 166, no. 1-2, pp. 43–62, 2002.

[45] L. Barnett, A. B. Barrett, and A. K. Seth, “Granger causality and transfer entropy Are equivalent for gaussian variables,” *Physical Review Letters*, vol. 103, no. 23, pp. 2–5, 2009.

[46] F. Takens, “Detecting strange attractors in turbulence,” in *Dynamical Systems and Turbulence* (D. Rand and L.-S. Young, eds.), pp. 366–381, Springer-Verlag, 1981.

[47] E. R. Deyle, M. C. Maher, R. D. Hernandez, S. Basu, and G. Sugihara, “Global environmental drivers of influenza.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 46, p. 201607747, 2016.

[48] E. B. Baskerville and S. Cobey, “Does influenza drive absolute humidity?,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. E2270–E2271, 2017.

[49] G. Sugihara, E. R. Deyle, and H. Ye, “Reply to Baskerville and Cobey: Misconceptions about causation with synchrony and seasonal drivers.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 12, pp. E2272–E2274, 2017.

[50] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu, “Methods for Quantifying the Causal Structure of Bivariate Time Series,” *International Journal of Bifurcation and Chaos*, vol. 17, no. 03, pp. 903–921, 2007.

[51] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, “Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance,” *Biological Cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[52] J. Pahle, A. K. Green, C. J. Dixon, and U. Kummer, “Information transfer in signaling pathways: a study using coupled simulated and experimental data.,” *BMC bioinformatics*, vol. 9, p. 139, 2008.

[53] G. V. Steeg and A. Galstyan, “Information Transfer in Social Media,” *Entropy*, vol. 90292, no. 1, pp. 1–8, 2011.

[54] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, “Local information transfer as a spatiotemporal filter for complex systems,” *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 77, no. 2, pp. 1–11, 2008.

[55] P. L. Williams and R. D. Beer, “Generalized measures of information transfer,” *arXiv:1102.1507*, pp. 1–6, 2011.