Supplementary Information for "On the predictability of infectious disease outbreaks"

Samuel V. Scarpino[1,2,†,*] **and Giovanni Petri**[3,†,**]

[1]Department of Mathematics & Statistics, University of Vermont, Burlington, VT, 05405, USA
[2]Complex Systems Center, University of Vermont, Burlington, VT, 05405, USA
[3]ISI Foundation, via Alassio 11c, Torino, 10126, Italy
[†]Both authors contributed equally to this work.
[*]svscarpi@uvm.edu
[**]giovanni.petri@isi.it

Additional data and code are available at `https://github.com/scarpino`

Methods

Permutation Entropy. In this Letter we make use of *permutation entropy* as a model-independent measure of the growth in complexity and unpredictability of infectious disease time series. Given a time series $\{x_t\}_{t=1,\ldots,N}$ indexed by positive integers, an embedding dimension d and a temporal delay τ , one can consider the set of all sequences of values s of the type $s = \{x_t, x_{t+\tau}, \ldots, x_{t+(d-1)\tau}\}$. Note that successive values $x_{t+i\tau}, x_{t+(i+1)\tau}$ for generic i can be in an arbitrary relative order. To each s, one can associates the permutation π of order d that makes s totally ordered, that is $\tilde{d} = \pi(d) = \{x_{t_i}, \ldots, x_{t_N}\}$ such that $x_{t_i} < x_{t_j} \forall t_i < t_j$. In this way, via π we associate a rank-order quantity that is independent of the actual values the timeseries takes and we can associate a probability p_π to each permutation by simply counting how many times if appears in the data as compared to the total number of sequences appearing. The permutation entropy of time series $\{x_t\}$ is then given by the Shannon entropy on the permutation orders, that is $H^p_{d,\tau}(\{x_t\}) = -\sum_\pi p_\pi \log p_\pi$. In the manuscript we show results obtained by fixing $\tau = 1$ to aid the intuition of the reader and select the most conservative (smallest) value of $H^p(\{x_t\}) = \min_d H^p_{d,\tau=1}(\{x_t\})$ by swiping over a wide range of possible d values. However, the qualitative results do not change even when we allow for a full swipe on (d, τ) pairs and setting $H^p(\{x_t\}) = \min_{d,\tau} H^p_{d,\tau}(\{x_t\})$.

Markov chain simulations

For each timeseries $\{x_t\}$, we build a Markov Chain defined on the corresponding codebook's alphabet, that is, we use the set of actual permutations explored by the system as state set for the Markov chain. For a time series with embedding dimension d we have thus $d!$ potential states. We then count the of transitions between two successive symbols along $\{x_t\}$ to define the transition probabilities between states of the Markov Chain. We then repeatedly generate series of symbols of the same length of the original timeseries, starting from a randomly chosen state, and finally and calculate the entropy of the state visits' distribution.

Literature R_0 estimates

Disease	Mean R_0	Range	Citation(s)
chlamydia	0.99	(0.43 – 1.49)	[1,2,3,4]
gonorrhea	1.34	(0.82 – 2.0)	[2,5,6]
hepatitis A	2.45	(0.40 – 4.0)	[7,8,9]
influenza	1.47	(0.9 – 2.1)	[10,11]
measles	15.10	(4.7 – 31.0)	[12,13,14,15]
mumps	9.94	(3.0 – 31.5)	[16,17]
polio	5.36	(4.0 – 7.0)	[12,18,19]
whooping cough	14.75	(5 – 20)	[18,20,12,21,22]
Zika	2.7	(0.50 – 6.3)	[23]

Table 1. Values of the basic reproductive ratio (R_0) for diseases included in this study were determined via literature review.

Significance tests on moving-window permutation entropy

We use a permutation test to determine whether different time series windows have distinct symbol distributions. Specifically, we fit a multinomial distribution to the normalized symbol frequency distributions and repeatedly simulate data from the estimated multinomials. Then, we calculate the Kullback-Leibler divergence between each pair of simulated distributions. With these simulated distributions, we can ask how often we see fluctuations in our estimate of the permutation entropy just due to sampling. More formally, we use these simulated distributions as a null distribution for calculating a frequentist p-value based on the observed

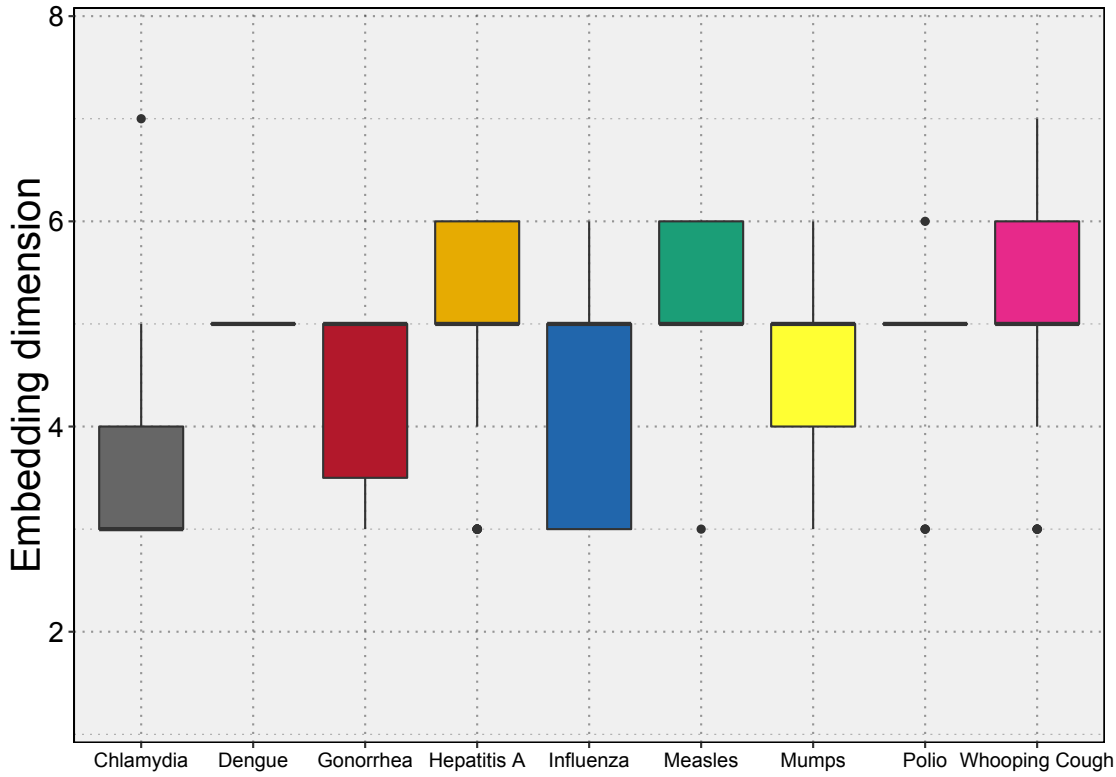

Figure S1. Embedding dimensions. We show the distributions of embedding dimensions d by disease. Embedding dimensions were obtained for each disease independently by minimizing the permutation entropy over a wide range of potential dimensions ($d \in [1, 20]$). Making this conservative choice on d allow us to interpret this length of the fundamental symbols used in computing permutation entropy as the natural temporal scale for the predictability of the corresponding timeseries. Notably, all diseases display narrow distributions and peaks between 3 and 6 weeks, with STDs, influenza and mumps being characterized by the shortest entropy horizons.

Kullback-Leibler divergence between the symbolic frequencies in time series windows.

References

1. Potterat, J. J. *et al.* Chlamydia transmission: concurrency, reproduction number, and the epidemic trajectory. *American Journal of Epidemiology* **150**, 1331–1339 (1999).

2. Brunham, R. C., Nagelkerke, N. J., Plummer, F. A. & Moses, S. Estimating the basic reproductive rates of Neisseria gonorrhoeae and Chlamydia trachomatis: the implications of acquired immunity. *Sexually transmitted diseases* **21**, 353–356 (1994).

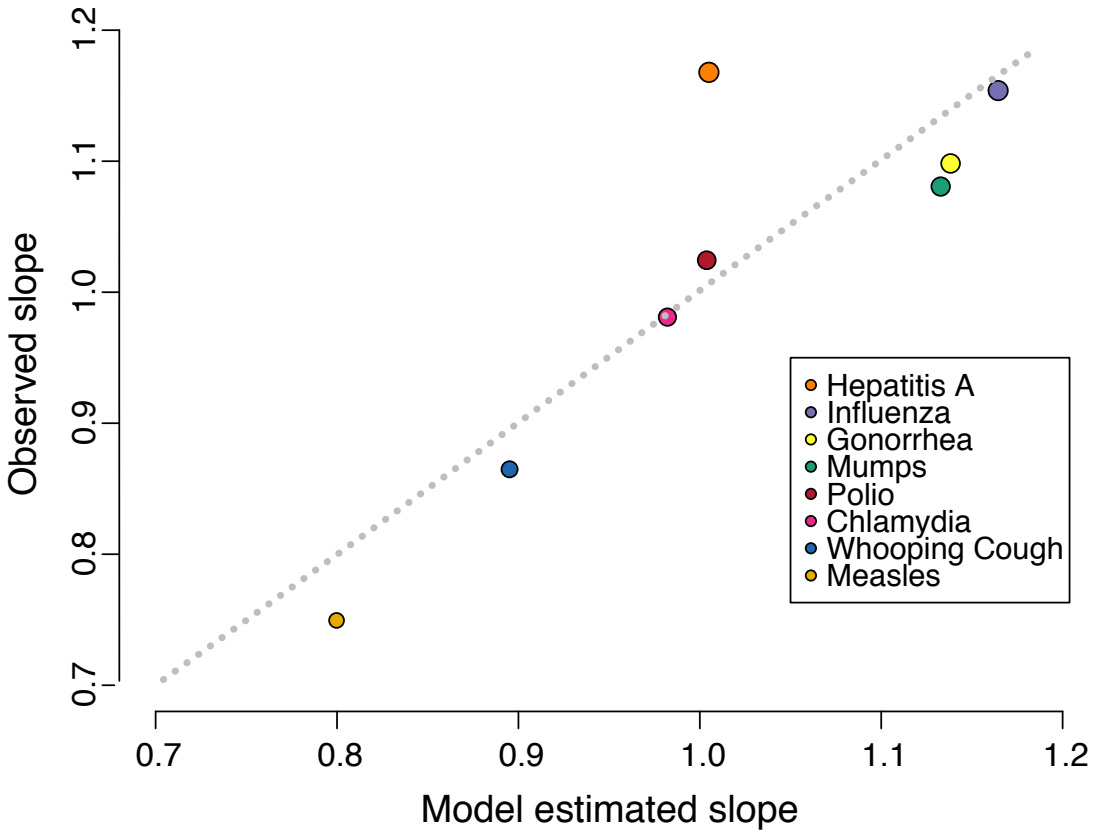

Figure S2. Slope of H^p growth. In the main text we show that a mixed effect model yields a linear relationship between the (log)entropy and (log)timeseries length, where disease has a random effect on the slope. We find that the disease specific slopes, i.e. the fixed effect slope plus the average random effect for each disease, can be predicted using only the embedding dimension, giving additional support to d as a fundamental dynamical feature of the underlying spreading process.

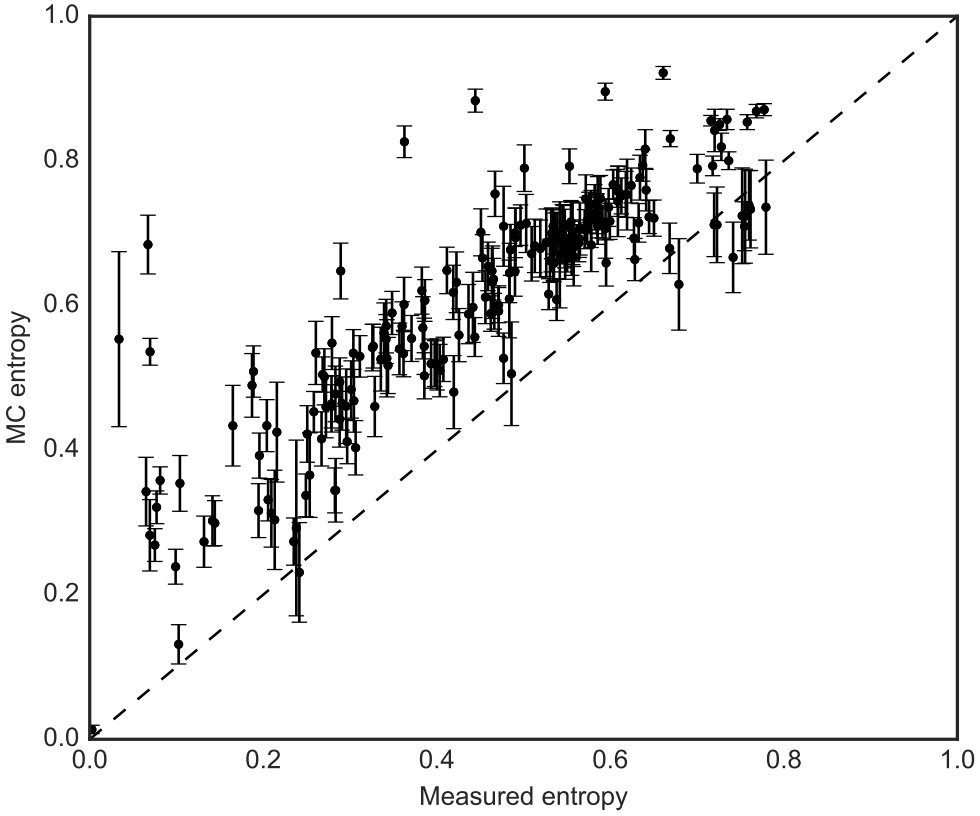

Figure S3. *d*-**th order Markov Chain entropy.** We simulated Markov Chains on the symbol codebooks extracted for each timeseries and we show that for all timeseries the estimated H^p is higher or comparable to the one computed from data. For clarity, we show here only the results for a selection of timeseries (with length between 300-1000 weeks) but the results apply to all series.

3. Althaus, C. L., Choisy, M. & Alizon, S. How sex acts scale with the number of sex partners: evidence from Chlamydia trachomatis data and implications for control. *PeerJ PrePrints* **3**, e1821 (2015).

4. Liu, F. *et al.* Assessment of transmission in trachoma programs over time suggests no short-term loss of immunity. *PLoS Negl Trop Dis* **7**, e2303 (2013).

5. McCluskey, C. C., Roth, E. & Van Den Driessche, P. Implication of Ariaal sexual mixing on gonorrhea. *American journal of human biology* **17**, 293–301 (2005).

6. Fingerhuth, S. M., Bonhoeffer, S., Low, N. & Althaus, C. L. Antibiotic-resistant Neisseria gonorrhoeae spread faster with more treatment, not more sexual partners. *PLoS Pathog* **12**, e1005611 (2016).

7. Regan, D. *et al.* Estimating the critical immunity threshold for preventing hepatitis A outbreaks in men who have sex with men. *Epidemiology and infection* **144**, 1528–1537 (2016).

8. Gay, N., Morgan-Capner, P., Wright, J., Farrington, C. & Miller, E. Age-specific antibody prevalence to hepatitis A in England: implications for disease control. *Epidemiology and infection* **113**, 113 (1994).

9. Van Effelterre, T. P., Zink, T. K., Hoet, B. J., Hausdorff, W. P. & Rosenthal, P. A mathematical model of hepatitis a transmission in the United States indicates value of universal childhood immunization. *Clinical infectious diseases* **43**, 158–164 (2006).

10. Pourbohloul, B. *et al.* Initial human transmission dynamics of the pandemic (H1N1) 2009 virus in North America. *Influenza and other respiratory viruses* **3**, 215–222 (2009).

11. Chowell, G., Miller, M. & Viboud, C. Seasonal influenza in the United States, France, and Australia: transmission and prospects for control. *Epidemiology and infection* **136**, 852–864 (2008).

12. Anderson, R. M., May, R. M. & Anderson, B. *Infectious diseases of humans: dynamics and control*, vol. 28 (Wiley Online Library, 1992).

13. Wichmann, O. *et al.* Large measles outbreak at a German public school, 2006. *The Pediatric infectious disease journal* **26**, 782–786 (2007).

14. van Boven, M. *et al.* Estimation of measles vaccine efficacy and critical vaccination coverage in a highly vaccinated population. *Journal of the Royal Society Interface* **7**, 1537–1544 (2010).

15. Grais, R. F. *et al.* Estimating transmission intensity for a measles epidemic in Niamey, Niger: lessons for intervention. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **100**, 867–873 (2006).

16. Whitaker, H. & Farrington, C. Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. *Statistics in medicine* **23**, 2429–2443 (2004).

17. Kanaan, M. & Farrington, C. Matrix models for childhood infections: a Bayesian approach with applications to rubella and mumps. *Epidemiology and Infection* **133**, 1009–1021 (2005).

18. Plotkin, S. A., Orenstein, W. A. & Offit, P. A. *Vaccines (Sixth Edition)* (W.B. Saunders, 2013).

19. Tebbens, R. J. D. & Thompson, K. M. Modeling the potential role of inactivated poliovirus vaccine to manage the risks of oral poliovirus vaccine cessation. *Journal of Infectious Diseases* **210**, S485–S497 (2014).

20. Kretzschmar, M., Teunis, P. F. & Pebody, R. G. Incidence and reproduction numbers of pertussis: estimates from serological and social contact data in five European countries. *PLoS Med* **7**, e1000291 (2010).

21. Althouse, B. M. & Scarpino, S. V. Asymptomatic transmission and the resurgence of Bordetella pertussis. *BMC medicine* **13**, 146 (2015).

22. de Cellès, M. D., Magpantay, F. M., King, A. A. & Rohani, P. The pertussis enigma: reconciling epidemiology, immunology and evolution. In *Proc. R. Soc. B*, vol. 283, 20152309 (The Royal Society, 2016).

23. Gao, D. *et al.* Prevention and control of Zika as a mosquito-borne and sexually transmitted disease: a mathematical modeling analysis. *Scientific reports* **6** (2016).