Patterns and scales in infectious disease surveillance data: an exploratory data analysis approach

Soetens, L.C.

4

Multi-scale analysis of infectious disease surveillance data

The contents of this chapter have been submitted for publication:
Multi-scale analysis of infectious disease surveillance data
Loes Soetens, Susan Hahné, Jacco Wallinga

Abstract

Surveillance systems collect information on many infections over long periods. To allow interpretation, these time series are summarized. Here we aim to identify the most informative summary statistics that capture dynamic patterns across scales in a multivariate time series of daily notifications for 45 infectious diseases from 2003 to 2018 in the Netherlands.

To put different flavors of summary statistics on the same footing, we calculate them as instances of the generalized Rényi entropy H_α, with H_0 corresponding to presence/absence, H_1 to epidemic intensity (Shannon entropy), and H_2 to the correlation sum. We then rank the order in which these statistics capture the dynamic variation across scales using Principal Component Analysis.

The Shannon entropy (H_1) is the most informative measure; the difference between presence/absence and correlation sum (H_0-H_2) is an informative orthogonal measure. We find a group of infections characterized by outbreaks, a group of infections continuously reported, and a group of infections intermittently reported. A few infections displayed a changing dynamic pattern: diphtheria and leptospirosis moved towards outbreaks, mumps and hantavirus moved away from outbreaks, and West Nile virus infection and yellow fever moved towards continuous reporting.

Our method for identifying changes in dynamic patterns of infectious diseases discerns disease groups sharing similar dynamic properties. By sorting diseases on recent change in dynamic properties, an informative overview is given of ongoing changes that might require enhanced disease control.

Introduction

Due to longstanding surveillance, databases containing case-based information on infectious disease notifications have expanded over the years. Epidemiologists use this information for outbreak detection, monitoring endemic disease trends, and the evaluation of intervention programs [1]. To achieve these aims, changes in dynamic patterns need to be identified; a change could be related to, for example, a starting outbreak, an increasing trend, or the introduction of an intervention program. To identify these changes, the number of disease notifications over time needs to be studied. This is often done separately for every disease, however, by doing so important relations between diseases dynamics might be missed. To allow for comparison of disease dynamics, they can be classified according to their dynamic properties. Dynamic properties can include the presence/absence of disease, and variation and autocorrelation in disease occurrence. When grouped, changes in these dynamic patterns might be more easily identified, enhancing infectious disease surveillance. The aim of this study is to provide a method for the identification of changes in dynamic patterns in multivariate infectious disease time series. To this end, we study the dynamic patterns of infectious diseases by using the complete database of the notifiable disease surveillance in the Netherlands, comprising 49 infectious diseases reported daily over sixteen years.

We use exploratory data visualisation to get an overview of large amounts of data [2]. So what measure can be used to visualize disease dynamics? The number of notifications can be visualized by plotting time series of incidence [3, 4]. However, the representation of such a time series is heavily dependent on the choice of the time aggregation scale. Other measures of disease dynamics discussed in literature concern measures to quantify the presence or absence of disease in a population [5-11]. These are often used to measure persistence of a pathogen in the population by counting either the number of or the total duration of extinctions in the population. The dynamics of measles have frequently served as a case study to investigate these patterns [6-9]. Another measure of disease dynamics is Shannon entropy [12]. Applied to infectious disease time series it can serve various purposes: as an early warning measure [13], to quantify epidemic intensity [14] or to forecast infectious disease incidence [15]. A final group of measures for capturing disease dynamics are measures of (auto)correlation, such as obtained with wavelet or spectral analysis. The power of these methods lies in their ability to describe how patterns change across aggregation scales [16]. This category has probably been the most frequently applied in the infectious disease domain [17-20].

All these methods often only cover one aspect of disease dynamics. To capture multiple aspects of disease dynamics, visualizing disease dynamics requires an easy to understand measure that is robust to aggregation scale and that can be applied to a large number of

infectious diseases. Here, we propose Rényi entropy as this measure [21]. Rényi entropy is a family of entropy measures that can be used to quantify diversity of phenomena, a so-called diversity index [22]. It is a generalisation of the earlier mentioned Shannon entropy. We chose this entropy family for various reasons. It is very straightforward to calculate; it only uses relative frequencies of notifications occurring at a specific timescale. We can therefore easily estimate the entropies along a continuum of aggregation levels. This has two advantages; we can examine how patterns change across these levels and separate members of the family all quantify another important aspect of time dynamics that may be related to measures mentioned in the previous paragraph, such as the presence/absence of notifications, uncertainty around the mean (Shannon entropy), and (auto)correlation. By calculating this single index, we can therefore capture multiple dimensions of disease dynamics.

We apply our approach to a large database with surveillance data of notifiable diseases in the Netherlands from 2003 to 2018, comprising 49 infectious diseases, ranging from frequently reported diseases, such as pertussis, to rare diseases, such as botulism. We visualize time series of all 49 diseases included in the database, and group and sort these time series according to their similarity. We study changes in the classification of diseases over time.

Methods

Data

We use all infectious disease notifications in the Netherlands between 2003-2018. The date of onset is used in our analysis; when this is missing we used date of laboratory confirmation; and if that was missing we used the date of diagnosis. We stratify the hepatitis B and C notifications into chronic and acute infections, and meningococcal infections into serogroups B, C, W and Y. If the hepatitis infection type (i.e., acute vs. chronic) or meningococcal serogroup is unknown or classified as 'other', the relevant cases are excluded. Cases with chronic hepatitis C infection are excluded as this infection was no longer notifiable from 2003. Cases with influenza A(H1N1)2009 infection are excluded as this infection was only notifiable for a short period of time (2009-2010). Dengue or chikungunya virus infections are excluded as these are only notifiable in the Dutch Caribbean areas. Only case-based notifications are included because a date of notification is needed for every record; this means exclusion of notifications of food-borne and MRSA clusters. Not all diseases were notifiable from the beginning of the study period (in 2003). For mumps, Hantavirus infection, Haemophilus influenzae type b (Hib) disease, pneumococcal disease, listeriosis and tuberculosis, cases are included from 2009 onwards; for group A streptococcal (GAS) disease cases are included from 2011 onwards; and for tularemia and zika virus infection, cases are included from 2017 onwards.

From time series to discrete frequency distributions

We consider time series (X) of notification counts. These time series can be aggregated at a level i (such as days, weeks, months or years). We can then calculate the relative frequencies p_i of notifications in the i^{th} time window, transforming our time series in discrete frequency distributions. For example, if the time series is aggregated by week and in week 4 there are 10 notifications and the total number of notifications is 100, the relative frequency in week 4 is p_4= 10/100= 0.1. We transform the time series to discrete frequency distributions for every disease, at various aggregation levels (as discussed below). The Rényi entropy family can then be described by the following formula [21]:

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} p_i^\alpha\right)$$

where n represents the total number of time windows in the time series. The order α is indicating the family member and can vary from zero to infinity.

H(X) based on each level of α can be interpreted in its own right as a measure corresponding to a real phenomenon. In this study, we use particular instances or family members of the Rényi entropy:

- α = 0 is quantifying the presence/absence of notifications. If all time windows

contain notifications this is equal to the maximum entropy (H_{max}), which is defined as the logarithm of the time series length.

- $\alpha \to 1$, better known as Shannon entropy, is quantifying the fluctuations in notifications. When there are no fluctuations in the time series, i.e. a similar amount of notifications in every time window, this measure equals 1. When there is only one time window with notifications, this measure equals zero.
- $\alpha = 2$ captures the probability that two notifications, taken at random from the time series, come from the same time window of length τ. Specifically, H_2 is the integral over the (auto)correlation function for the time series, taken from 0 to τ. As such, it quantifies the autocorrelation in the occurrence of notifications. Again, for time series with a constant number of notifications throughout the series this measure will equal the maximum entropy.
- $\alpha \to \infty$, the entropy then equals the maximum p_i value in the time series, the proportional frequency of the most frequent time window. This is also called the minimum entropy.

The values of the Rényi entropy of order α, H_α, are ordered as $H_{max} \geq H_0 \geq H_1 \geq H_2 \geq H_{min}$ [21].

Analysis

Overview of the time series

To visualize the time series of diseases, we plot the aggregated number of notifications as a heat map, where rows represent diseases, and columns represent time windows. Diseases are ordered alphabetically. A color scheme indicates the number of notifications per time window. We map this scheme on a log scale as there is large variation in the incidence. Time windows without cases (log(0)) are left blank. We construct a heat map per week and per year, to show differences in patterns when aggregation scales vary. Second, we calculate the Rényi entropies of order 0, 1 and 2 and the maximum and minimum entropy based on the time series of every disease for various aggregation levels over the complete inclusion range. We aggregate the time series by the following time window sizes: day, week, month, 3 months, 6 months, 9 months, year, 1.5 years, 2 years, 3 years and 5 years. We plot the results in a multi panel plot, in which the values of the entropies are depicted against the natural logarithm of the aggregation level (in days) for every infectious disease.

Grouping of infectious diseases

To identify groups of infectious diseases with similar dynamic patterns, we perform principal component analysis on the Rényi entropies of order 0, 1 and 2, on the complete time series of all diseases at various aggregation levels. As principal component analysis is applied to only three entropy measures, we are able to interpret these components by examining the contributions of the individual entropies to the principal components (PC) across the various aggregation levels. This is unusual in principal component analysis, which often involves

large numbers of variables, making the interpretation of the components impossible. The change of the contributions across aggregation levels will then give us an assessment of the robustness of the entropy values across aggregation levels. The results of the principal component analysis are plotted in a scatterplot per aggregation level to assess which diseases are closest to each other and therefore share similar time dynamics.

Change in dynamic patterns over time

To examine the change in dynamic patterns for the various diseases, entropies of order 0, 1 and 2 are calculated per disease over three-year moving windows, shifting by year. The aggregation level is based on the previous analyses. We use three-year periods to be able to take into account multiyear outbreaks and seasonal effects. The results are plotted in a graph with the interpretation of the PC in terms of entropies on the x- and y-axis and a path through time per disease, showing the change in dynamic patterns over time. Diseases with considerable dynamic change in their patterns will therefore show much variation and hence a long path, while for relatively stable diseases the path will be much shorter. Finally, we compare the dynamic change in the last two time periods (t_1= 2015-2017 versus t_2= 2016-2018) between diseases, to assess most recent dynamics. We define dynamic change as the Euclidian distance in entropies between (x_{t1}, y_{t1}) and (x_{t2}, y_{t2}) (the length of the path between these time periods). We sort diseases in the heatmap of disease incidence according to dynamic change in four different panels: one for disease incidence, one showing the decreasing dynamic change, and two with the x- and y- components (the interpretation of the PC in terms of entropies) to help with the interpretation of the dynamic change. In Appendix B, we repeat this analysis for two different time periods.

Results

In the period 2003 – 2018, 210,715 infectious disease cases were notified in the Netherlands. 196,779 cases were included, distributed over 45 infectious diseases (Table 4.1). Pertussis cases represent the vast majority (56.6%) of included cases.

In Figure 4.1a the time series of weekly notifications per infectious disease are shown on a log10 color scale. We observe several diseases with large outbreaks in the past 15 years, such as measles, rubella and Q fever; diseases with a stable number of cases, such as pertussis and chronic hepatitis B; diseases with a more irregular pattern, such as hepatitis A; diseases with a strong seasonal component, such as legionellosis; and very rare diseases, such as trichinosis, rabies and botulism. In Figure 4.1b the same time series are aggregated per year. Here, seasonal patterns are lost.

For every infectious disease we have calculated Rényi entropies of order 0, 1 and 2 and the maximum and minimum entropy based on time series at various aggregation levels (Figure 4.2). At a daily aggregation level, a wide gap exists between the curves of H_{max} and H_0 for diseases such as brucellosis (BRUC), rubella (RUBE) and meningococcal C, W and Y disease (MCOCC, MCOCW, MCOCY). This means that these diseases have many days without reported cases. At a 5-year aggregation level, we can only observe a small gap between these curves for very rare diseases, such as rabies (RABI) and anthrax (ANTH). We can also observe diseases, such as pertussis (PERT), tuberculosis (TBC), chronic hepatitis B (HEPBCH) and shigella (SHIG), for which all orders of the Rényi entropies are close to the H_{max} across all aggregation levels. These are diseases with a continuous reporting of cases (H_0), with little fluctuations around the mean number of reported cases (H_1) and low autocorrelation in the number of reported cases (H_2). Finally, we have a group of diseases, such as measles (MEAS), Q fever (QFEV), rubella (RUBE) and meningococcal W disease (MCOCW), which show a relatively large gap between the curves of order 1 and 2 and the curve of order 0, especially at monthly and yearly aggregation levels. These are diseases for which if cases are reported, large fluctuations around the mean number of reported cases (H_1) and a large variation in the number of reported cases throughout the time series are shown. We defined this as a typical pattern for outbreak diseases.

Table 4.1. Overview of notifiable disease cases 2003 – 2018 included in this study

Disease	Disease short	Start year	Number of notifications
Anthrax	ANTH	2003	2
Botulism	BOTU	2003	21
Brucellosis	BRUC	2003	81
Cholera	CHOL	2003	39
Creutzfeldt-Jakob disease	CJD	2003	345
Variant Creutzfeldt-Jakob disease	CJDV	2003	3
Diphtheria	DIPH	2003	16
Hantavirus infection	HANTA	2009	239
Hepatitis A	HEPA	2003	3,717
Hepatitis B acute	HEPBAC	2003	3,489
Hepatitis B chronic	HEPBCHR	2003	21,679
Hepatitis C acute	HEPCAC	2003	846
Invasive group A streptococcal disease	IGAS	2011	1,610
Invasive Haemophilus influenzae type b infection	IHIB	2009	253
Human infection with zoonotic influenza virus	INFL	2009	1
Invasive pneumococcal disease (in children 5 years or younger)	IPNEU	2009	456
Legionellosis	LEGI	2003	6,875
Leptospirosis	LEPT	2003	843
Listeriosis	LIST	2009	1,152
Malaria	MALA	2003	4,360
Meningococcal disease, serogroup B	MCOCB	2003	2,233
Meningococcal disease, serogroup C	MCOCC	2003	158
Meningococcal disease, serogroup W	MCOCW	2003	276
Meningococcal disease, serogroup Y	MCOCY	2003	168
Measles	MEAS	2003	3,246
Mumps	MUMPS	2009	2,181
Paratyphoid A fever	PARA	2003	263
Paratyphoid B fever	PARB	2003	351
Paratyphoid C fever	PARC	2003	24
Pertussis	PERT	2003	111,360
Psittacosis	PSIT	2003	1,018
Q fever	QFEV	2003	4,704
Rabies	RABI	2003	4
Rubella	RUBE	2003	595
Shigella	SHIG	2003	7,795
STEC/enterohemorragic E.coli infection	STEC	2003	6,605
Tuberculosis	TBC	2009	9,298
Tetanus	TETA	2009	15
Trichinosis	TRIC	2003	3
Tularemia	TUL	2017	5
Typhoid fever	TYPH	2003	427
Viral hemorrhagic fever	VHF	2003	2
West Nile virus infection	WNV	2009	5
Yellow fever	YFV	2003	3
Zika virus infection	ZIKA	2017	13
Total			**196,779**

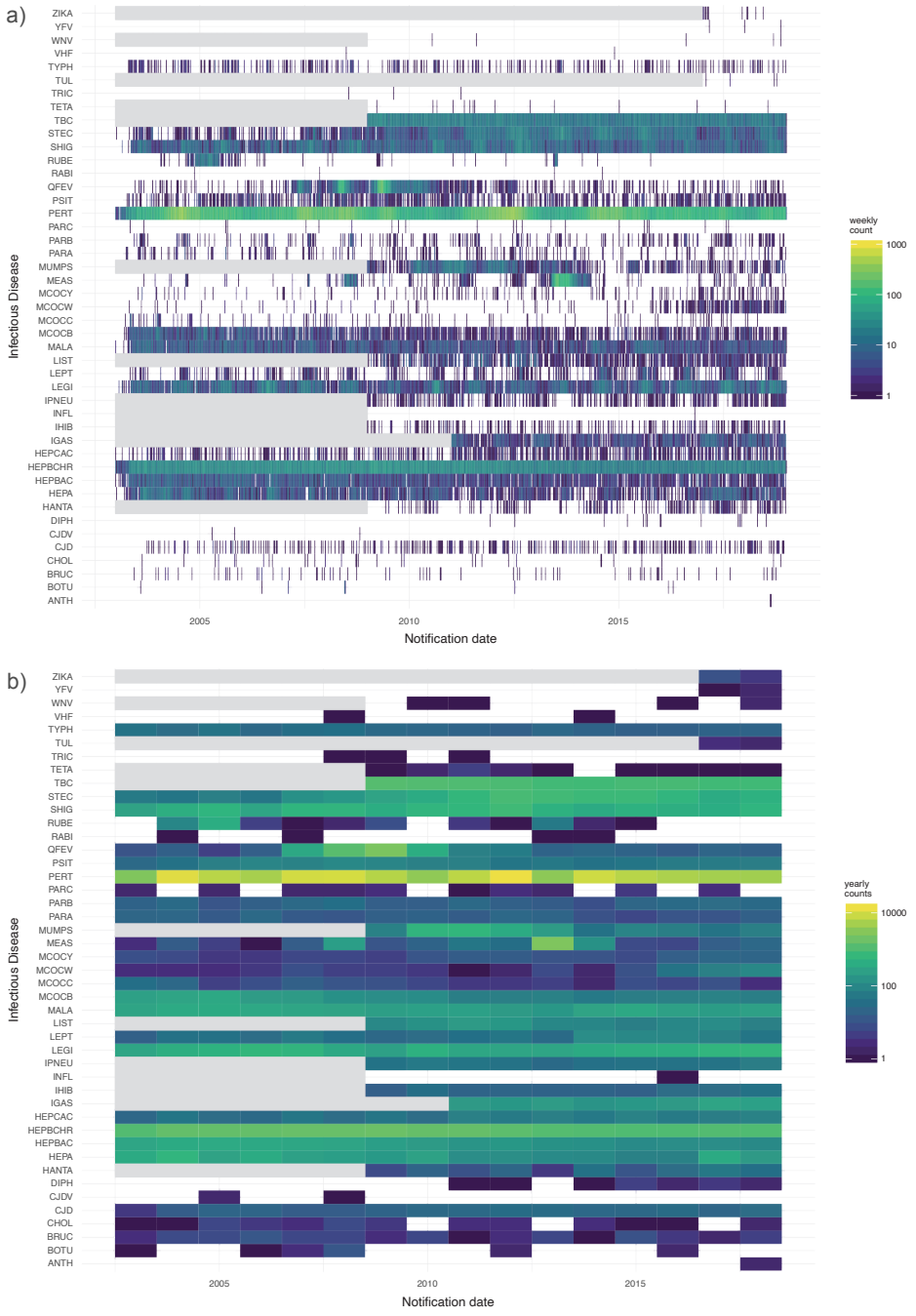

Figure 4.1. A first visualization of the notification data for 45 infectious diseases in the Netherlands, 2003 – 2018, aggregated by week (a) and by year (b). Grey areas indicate periods in which diseases were not yet notifiable.

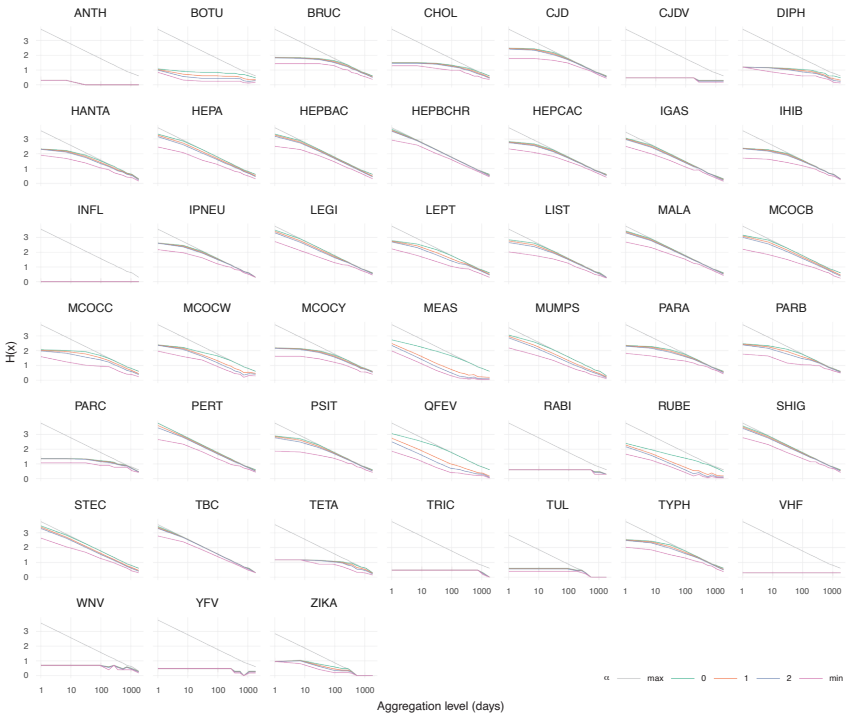

Figure 4.2. Rényi entropy for different orders of α against the aggregation level in days of the time series of notifications of 45 infectious diseases in the Netherlands, 2003-2018.

Principal Component Analysis reveals that dynamic patterns of infectious diseases can be characterized by just two variables, of which the most important variable is the intensity of fluctuations (H_1), and the lesser important one the sensitivity of this intensity to a change in order α (H_0-H_2) (Appendix A). Together, these two variables explain almost 100% of the variance in H_0, H_1 and H_2 of the diseases. Three groups of diseases are identified based on these properties: those characterized by outbreaks (high H_0-H_2, low H_1), those which are continuously reported (low H_0-H_2, high H_1), and those which are intermittently reported (low H_0-H_2, low H_1).

To examine the change in dynamics over time, entropies are calculated for different periods (Figure 4.3). Of interest are diseases that were formerly rare but over time are more constantly reported (shifting from bottom left to bottom right corner in Figure 4.3). Examples of such diseases are West Nile virus infection (WNV) or yellow fever (YFV). Also, diseases moving in the other direction on the same axis can be of interest: from being continuously reported to more fluctuations in their time series. This always goes together with movement along the other axis and indicates a group of diseases that show more fluctuations, less autocorrelation and few intervals with no reported cases in their time

series over the years. These patterns might possibly indicate (starting) outbreaks (diseases that shift from the bottom upwards in Figure 4.3). Examples of such diseases are hepatitis A (HEPA), leptospirosis (LEPT), and diphtheria (DIPH). A final group of diseases are those with a large range in Figure 3. These diseases show large fluctuations over time, and are known for large past outbreaks. Examples are rubella (RUBE), measles (MEAS) and Q fever (QFEV). Diseases with constant dynamics over time include acute and chronic hepatitis B infection (HEPBAC and HEPBCHR), tuberculosis (TBC), psittacosis (PSIT) and meningococcal B infection (MCOCB).

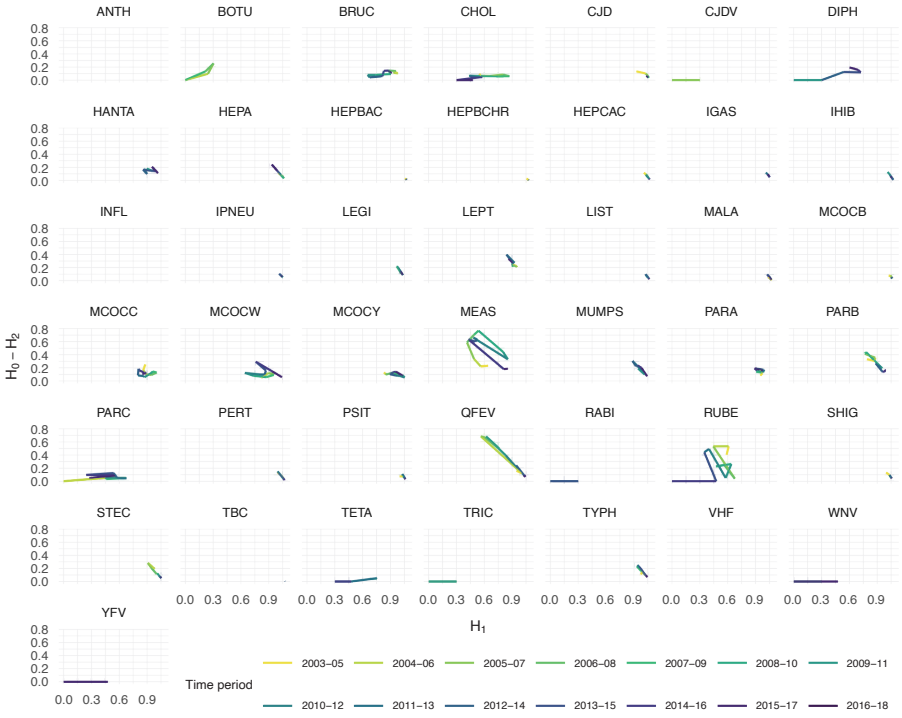

Figure 4.3. Change in entropy dynamics over time per infectious disease. A rolling window of three years, shifting per year, was applied on quarterly aggregated time series.

To get an overview of recent changes in dynamic patterns, we zoom in on the last two time periods presented in Figure 4.3 (purple). We can compare the change in dynamics between 2015-2017 and 2016-2018 by computing dynamic distance (the length of the path in Figure 4.3) between these periods and sort the heat map according to this distance (Figure 4.4). The largest change in dynamics between these periods was for West Nile virus infection (WNV) and yellow fever virus infection (YFV). This is due to an increase in H_1 entropy (see right panels of Figure 4.4). Previously, we observed that diseases with a decreasing H_1 and increasing H_0-H_2 difference possibly indicate starting outbreaks. We observe this pattern also in diphtheria (DIPH), Creutzfeldt Jakob disease (CJD) and leptospirosis

(LEPT). We observe an opposite pattern, a decrease in outbreak potential (decrease in H_0-H_2 distance, increase in H_1 distance), for mumps, meningococcal W disease (MCOCW), hantavirus infection (HANTA), hepatitis A (HEPA) and meningococcal Y disease (MCOCY) (amongst others). For most diseases, little change in dynamics has taken place between the periods 2015-2017 and 2016-2018. This is completely different when we compare two other periods, 2010-2012 and 2011-2013 (Figure B1, Appendix B). Here we can see large changes in dynamics for measles (MEAS) and rubella (RUBE), due to an increase in outbreak potential (H_0-H_2 distance) and a decrease in H_1 entropy, indicating more fluctuating time series.

Figure 4.4. Change in disease dynamics according to entropy measures between period t_1 (2015-2017) and t_2 (2016-2018) across notifiable infectious diseases in the Netherlands. The quarterly aggregated time series of these periods shown as a heat map (first panel) are sorted by decreasing dynamic distance between these periods (second panel).The third and fourth panel show the distance in H_0-H_2 and distance in H_1 between t_1 and t_2 for all diseases. Light grey areas indicate diseases for which no data was reported as the disease was not yet notifiable (first panel) or that the entropies could not be estimated due to no reported cases in one or both periods (second, third and fourth panel).

Discussion

We captured the intrinsic dynamic properties of 45 notifiable infectious diseases between 2003 and 2018 in the Netherlands, using a measure based on the relative incidence of infectious diseases. We have used Rényi entropy, a measure for which the outcome is consistent across timescales, to classify and order diseases by their dynamic properties.

Principal Component Analysis identified the most important quantity describing variation in dynamic patterns of diseases: H_1 (Shannon entropy), which gives a measure of intensity of fluctuations in the time series, and H_0-H_2, describing the difference in the entropy values, indicating the sensitivity of the intensity in fluctuations to a change in order.

Our aim was to provide a method for the identification of changes in dynamic patterns in multivariate infectious disease time series. Recent changes requesting further analysis by infectious disease epidemiologists include the more continuously reporting of West Nile virus and the more fluctuating reporting rate of diphtheria notifications in the Netherlands in 2018. In 2018, the incidence of West Nile virus infection was relatively high in South European countries [23-25], causing more imported cases in the Netherlands. Diphtheria was acquired abroad. Traveling or immigration from high incidence countries, such as Venezuela [26] (largest neighbouring country of the Kingdom of the Netherlands) or Indonesia [27] (former Dutch colony), might explain the more fluctuating pattern for this disease.

In this study, we have shown that Rényi entropy can be calculated at multiple aggregation scales and that the ranking of diseases on entropy measures is quite similar across different aggregation levels. The analysis does therefore not require a choice on the timescale of analysis, which is a major advantage in the case that no aggregation level is a priori of interest. Another advantage of this measure is that it is very easy to calculate, giving a low computational burden even for large datasets. As it relies on relative frequencies, diseases with low numbers can also be analysed. A change in the dynamics in the notifications of diphtheria in 2018 could therefore be observed, an issue that would otherwise probably have gone unnoticed.

So how can the visualisation of Rényi entropy be used in daily practice? We would advise to use this alongside regular surveillance activities, to check on a regular basis whether any important changes have occurred in disease dynamics. Hidden intrinsic patterns, that currently go unnoticed in regular surveillance activities (for example with rare diseases) can then be revealed and further investigated. Entropy might also be useful as a measure for early warning of outbreaks as suggested by Brett et al [13]. Our research has additionally demonstrated that change in patterns can be detected, but whether this can be done in a timely manner, as is requested for early warning, needs further research.

This method can be applied to any infectious disease surveillance data, as many developed countries have similar surveillance systems yielding similar data [3, 19, 28, 29], it would also be interesting to compare grouping of diseases according to their dynamic properties across countries. If similar patterns exist, new hypotheses on disease dynamics might be generated and typical patters for diseases might be revealed, as in Graham et al [30]. Finally, the analysis can also take into account age, sex, specific risk groups, or geography, as was done by Dalziel et al [14], to reveal and compare dynamic patterns between these subgroups.

To conclude, we have introduced a multi-scale measure for representing the dynamic properties of infectious diseases. By applying this measure on notifications of 45 infectious diseases in the Netherlands from 2003 -2018, we were able to discern disease groups sharing similar dynamic properties. In addition, by sorting diseases by the amount of change in dynamic properties, an informative overview is given of ongoing changes that might require the attention of epidemiologists and allow for the generation of hypotheses on disease dynamics.

References

1. Giesecke J. Modern Infectious Disease Epidemiology, Third Edition: Taylor & Francis; 2016.
2. Tukey JW. Exploratory Data Analysis: Addison-Wesley Publishing Company; 1977.
3. van Panhuis WG, Grefenstette J, Jung SY, Chok NS, Cross A, Eng H, et al. Contagious diseases in the United States from 1888 to the present. The New England journal of medicine. 2013;369(22):2152-8.
4. Gibney KB, Cheng AC, Hall R, Leder K. An overview of the epidemiology of notifiable infectious diseases in Australia, 1991-2011. Epidemiology and infection. 2016;144(15):3263-77.
5. Bjornstad ON, Grenfell BT. Noisy clockwork: time series analysis of population fluctuations in animals. Science (New York, NY). 2001;293(5530):638-43.
6. Keeling MJ, Grenfell BT. Understanding the persistence of measles: reconciling theory, simulation and observation. Proceedings Biological sciences. 2002;269(1489):335-43.
7. Bartlett MS. Measles Periodicity and Community Size. Journal of the Royal Statistical Society Series A (General). 1957;120(1):48-70.
8. Grenfell BT, Bjørnstad ON, Finkenstädt BF. DYNAMICS OF MEASLES EPIDEMICS: SCALING NOISE, DETERMINISM, AND PREDICTABILITY WITH THE TSIR MODEL. 2002;72(2):185-202.
9. Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model. Ecological Monographs. 2002;72(2):169-84.
10. Nieddu GT, Billings L, Kaufman JH, Forgoston E, Bianco S. Extinction pathways and outbreak vulnerability in a stochastic Ebola model. Journal of the Royal Society, Interface. 2017;14(127).
11. Vredenburg VT, Knapp RA, Tunstall TS, Briggs CJ. Dynamics of an emerging disease drive large-scale amphibian population extinctions. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(21):9689-94.
12. Shannon CE. A Mathematical Theory of Communication. 1948;27(3):379-423.
13. Brett TS, Drake JM, Rohani P. Anticipating the emergence of infectious diseases. Journal of the Royal Society, Interface. 2017;14(132).
14. Dalziel BD, Kissler S, Gog JR, Viboud C, Bjornstad ON, Metcalf CJE, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science (New York, NY). 2018;362(6410):75-9.
15. Scarpino SV, Petri G. On the predictability of infectious disease outbreaks. Nature communications. 2019;10(1):898.
16. Levin SA. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. 1992;73(6):1943-67.
17. Holdsworth AM, Kevlahan NK, Earn DJ. Multifractal signatures of infectious diseases. Journal of the Royal Society, Interface. 2012;9(74):2167-80.
18. van Wijhe M, Tulen AD, Korthals Altes H, McDonald SA, de Melker HE, Postma MJ, et al. Quantifying the impact of mass vaccination programmes on notified cases in the Netherlands. Epidemiology and infection. 2018;146(6):716-22.
19. Grenfell BT, Bjørnstad ON, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature. 2001;414(6865):716-23.
20. Soetens LC, Boshuizen HC, Korthals Altes H. Contribution of seasonality in transmission of Mycobacterium tuberculosis to seasonality in tuberculosis disease: a simulation study. American journal of epidemiology. 2013;178(8):1281-8.
21. Renyi A, editor On Measures of Entropy and Information. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics; 1961 1961; Berkeley, Calif.: University of California Press.
22. Tóthmérész B. Comparison of different methods for diversity ordering. 1995;6(2):283-90.
23. Burki T. Increase of West Nile virus cases in Europe for 2018. The Lancet. 2018;392(10152):1000.
24. Holt E. West Nile virus spreads in Europe. The Lancet Infectious Diseases. 2018;18(11):1184.
25. Haussig JM, Young JJ, Gossner CM, Mezei E, Bella A, Sirbu A, et al. Early start of the West Nile fever transmission season 2018 in Europe. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2018;23(32).
26. Paniz-Mondolfi AE, Tami A, Grillet ME, Márquez M, Hernández-Villena J, Escalona-Rodríguez MA, et al. Resurgence of Vaccine-Preventable Diseases in Venezuela as a Regional Public Health Threat in the Americas. Emerging infectious diseases. 2019;25(4):625-32.

27. Tosepu R, Gunawan J, Effendy DS, Ahmad LOAI, Farzan A. The outbreak of diphtheria in Indonesia. Pan Afr Med J. 2018;31:249-.
28. Enki DG, Garthwaite PH, Farrington CP, Noufaily A, Andrews NJ, Charlett A. Comparison of Statistical Algorithms for the Detection of Infectious Disease Outbreaks in Large Multiple Surveillance Systems. PloS one. 2016;11(8):e0160759.
29. Enki DG, Noufaily A, Garthwaite PH, Andrews NJ, Charlett A, Lane C, et al. Automated biosurveillance data from England and Wales, 1991-2011. Emerging infectious diseases. 2013;19(1):35-42.
30. Graham M, Winter AK, Ferrari M, Grenfell B, Moss WJ, Azman AS, et al. Measles and the canonical path to elimination. 2019;364(6440):584-7.

Appendix A

PRINCIPAL COMPONENT ANALYSIS RESULTS ACROSS AGGREGATION LEVELS

To identify groups of infectious diseases with similar dynamic patterns, we perform principal component analysis on the Rényi entropies of order 0, 1 and 2, based on the complete time series of all diseases at various aggregation levels. As principal component analysis is applied to only three entropy measures, we are able to interpret these components by examining the contributions of the individual entropies to the principal components (PC) across the various aggregation levels. This is unusual in principal component analysis, which often involves large numbers of variables, making the interpretation of the components impossible. The change of the contributions across aggregation levels will then give us an assessment of the robustness of the entropy values across aggregation levels. The results of the principal component analysis are plotted in a scatterplot per aggregation level to assess which diseases are closest to each other and therefore share similar time dynamics. It can be seen in Table A1 and Figure A1 that the contribution of individual entropies to the three principal components is very consistent across time scales: H_0, H_1, and H_2 contribute equally to principal component 1 (PC1), H_0 for 60% and H_2 for 40% to principal component 2 (PC2) and H_1 for 60% and H_2 for 40% to principal component 3 (PC3). Together PC1 and PC2 explain almost 100% of the variance in H_0, H_1, and H_2 of the diseases. Based on these contributions and since we only have three quantities in the analysis, we can interpret the principal components: PC1 can be interpreted as the mean of H_0, H_1, and H_2. As $H_0 \geq H_1 \geq H_2$, the mean of these entropies will be almost similar to H_1, the Shannon entropy. This justifies the use of Shannon entropy to characterize dynamic patterns by e.g. Dalziel et al [14]. PC2 can be interpreted as the difference between H_0 and H_2, describing the variance in entropies. Substituting principal components with the entropy terms yields similar figures (Figure A1 and A2). This indicates that the variation between diseases can be explained by (a combination of) the individual entropies and that this is robust across different aggregation levels of the time series.

In Figure A2 the results of the principal component analysis are shown for different aggregation levels. On the x-axis H_1 is depicted, the approximation of principal component 1, and on the y-axis, the difference between H_0 and H_2 is shown, the approximation of principal component 2. A considerable proportion of infectious diseases cluster together in the bottom right corner, indicating diseases with a high mean of the entropies (H_1) and low H_0-H_2 . These are continuously reported diseases with little fluctuations in their time series. Examples of such diseases are tuberculosis (TBC) and chronic hepatitis B (HEPBCHR). If we move from the bottom right corner to the bottom left corner, we move from diseases with little fluctuations in their time series, but which are less continuously reported, to the most extreme situation with one reported case ever of human infection with zoonotic influenza virus (INFL). If we move from the bottom right corner upwards, we come across

diseases with larger fluctuations in their time series, such as mumps, rubella, measles and Q fever, indicating outbreaks. The ratios between diseases across aggregation levels do not differ too much: with increasing aggregation level, the variation in H_1 between diseases decreases and the variation between diseases in H_0-H_2 peaks at the 3-month aggregation level after which it decreases again (Figure A1). It can also be observed that with increasing aggregation level, diseases are arranged on parallel diagonals. These diagonals represent periodicity. When considering the five year aggregation level, ZIKA and INFL are on the diagonal of one period of reported data, and for example, MCOCW and HEPBAC on the diagonal of four (all) periods of reported data.

Table A1. Relative contributions of H_0, H_1 and H_2 to the principal components at various aggregation levels.

		Contribution (%)		
Aggregation level	**Alpha**	**PC1**	**PC2**	**PC3**
day	0	35.42	53.61	10.97
	1	33.31	1.05	65.63
	2	31.27	45.34	23.39
week	0	35.75	58.23	6.02
	1	33.18	3.85	62.98
	2	31.07	37.93	31.00
month	0	35.41	60.33	4.27
	1	33.33	5.49	61.18
	2	31.26	34.19	34.55
3 months	0	33.88	62.54	3.58
	1	33.73	5.83	60.44
	2	32.39	31.62	35.98
6 months	0	31.97	64.82	3.21
	1	34.31	5.72	59.98
	2	33.72	29.47	36.81
9 months	0	31.67	65.60	2.73
	1	34.44	6.28	59.27
	2	33.89	28.12	38.00
year	0	30.61	66.83	2.56
	1	34.74	6.17	59.09
	2	34.65	27.00	38.35
18 months	0	31.50	65.78	2.72
	1	34.58	6.28	59.14
	2	33.92	27.94	38.14
2 years	0	31.37	65.81	2.82
	1	34.60	6.09	59.32
	2	34.03	28.10	37.86
3 years	0	31.72	65.93	2.36
	1	34.64	6.95	58.40
	2	33.64	27.12	39.24
5 years	0	38.40	59.03	2.57
	1	32.57	9.06	58.37
	2	29.02	31.91	39.07

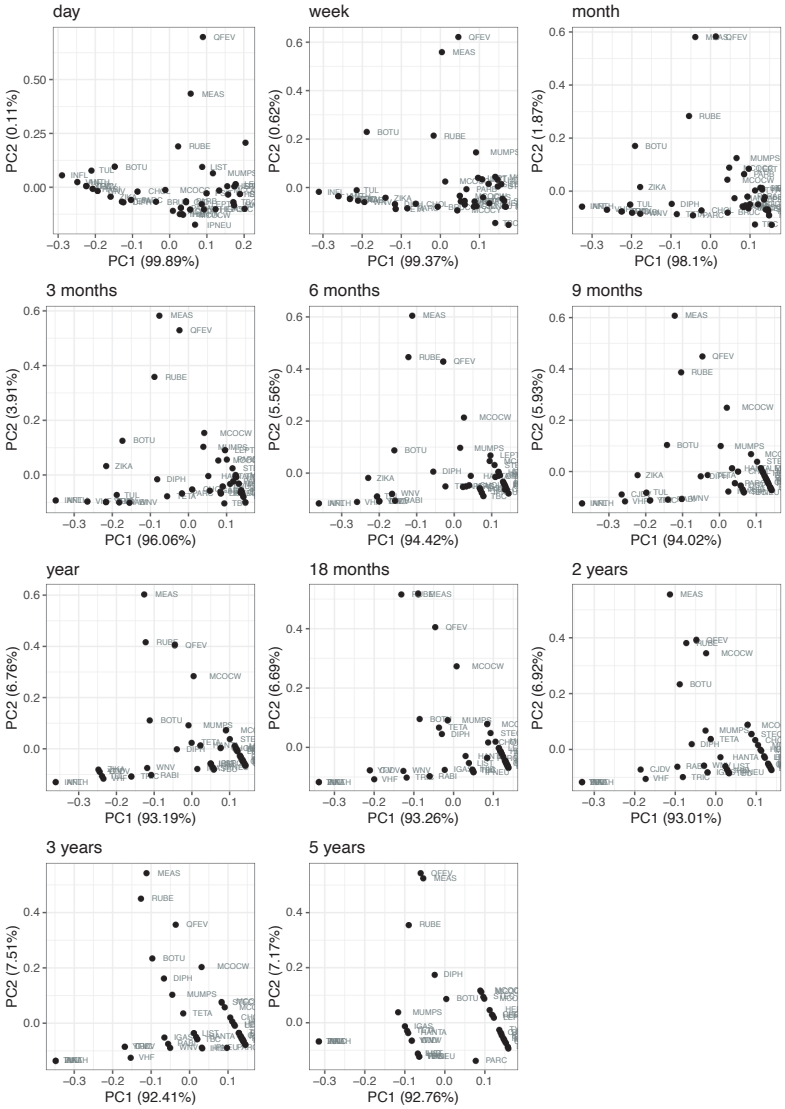

Figure A1. Results of the principal component analysis on H_0, H_1 and H_2 at various aggregation levels.

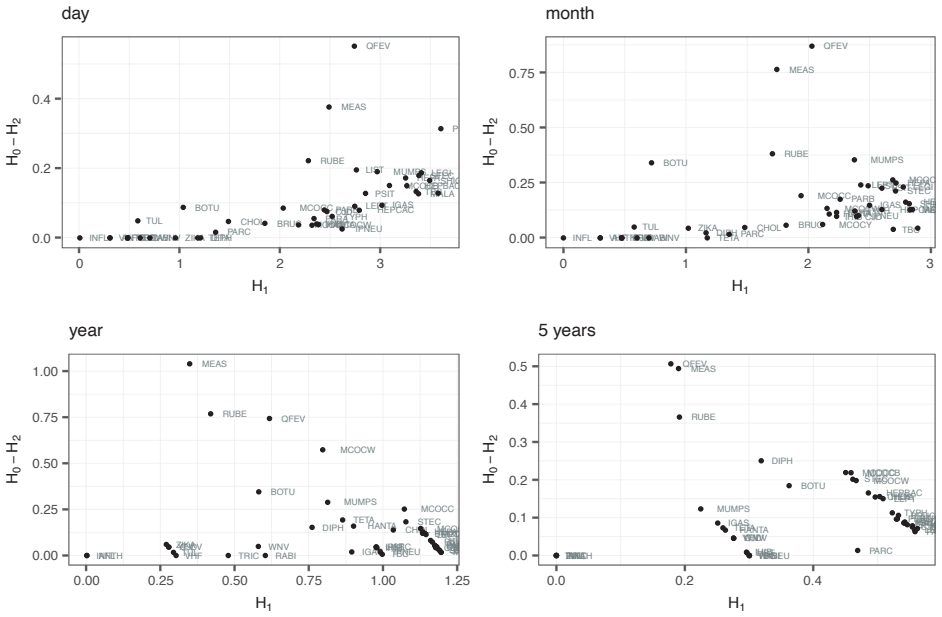

Figure A2. Results of the principal component analysis of H_0, H_1 and H_2 across different aggregation levels and for all diseases.

Appendix B

CHANGE IN DISEASE DYNAMICS ACCORDING ENTROPY MEASURES BETWEEN PERIOD T_1 (2010-2012) AND T_2 (2011-2013)

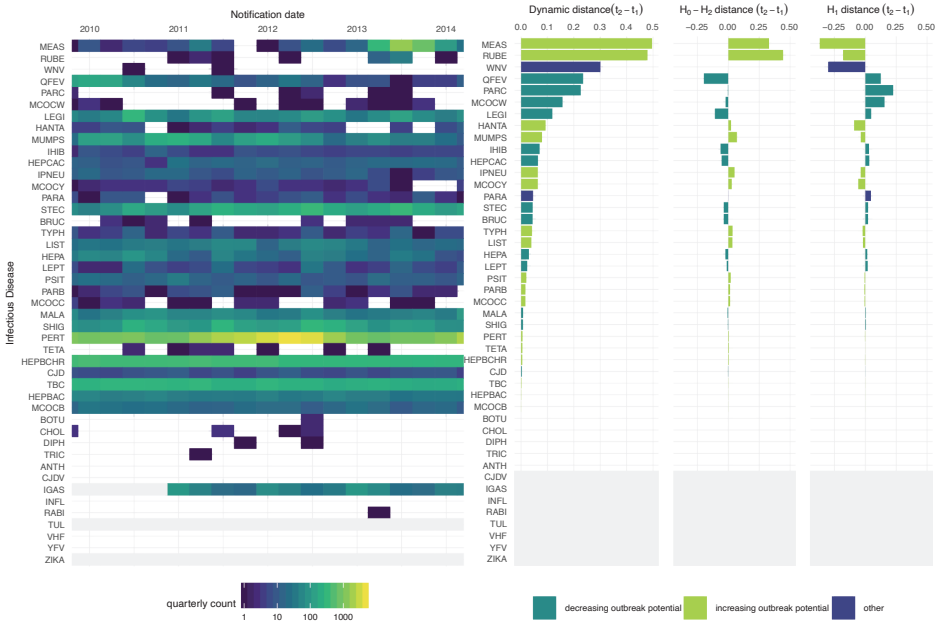

Figure B1. Change in disease dynamics according entropy measures between period t_1 (2010-2012) and t_2 (2011-2013) for all notifiable infectious diseases in the Netherlands. The quarterly aggregated time series of these periods shown as a heat map (first panel) are sorted by decreasing dynamic distance between these periods (second panel).The third and fourth panel show the distance in H_0-H_2 and distance in H_1 between t_1 and t_2 for all diseases. Light grey areas indicate diseases for which no data was reported as the disease was not yet notifiable (first panel) or that the entropies could not be estimated due to no reported cases in one or both periods (second, third and fourth panel).