Original Article

Scale-Free Distribution of Local Infection Clusters of Measles, Rubella, Syphilis, and HIV/AIDS: Correlation with Scale-Free Distribution of Municipality Population by Size

Hiroshi Yoshikura^{1*} and Fumihiko Takeuchi²

¹National Institute of Infectious Diseases, Tokyo 162–8640; and ²Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Tokyo 162–8655, Japan

SUMMARY: The size distribution of a local infection cluster (LIC), a group of infections reported from the same prefecture without interruption in successive weeks, was scale-free for infections that are transmitted from person-to-person (e.g., measles, rubella, syphilis, and HIV/AIDS). For infections that do not spread from person-to-person, the distribution was entirely random. The size distribution for measles, rubella, syphilis, and HIV/AIDS could be simulated successfully by random coin tossing with probabilities that were higher for highly populated prefectures. The size distribution of the population in large municipalities (> 120,000), as well as that of LICs, was found to be scale free. As the number of patients per prefecture was correlated with the equation $P = kN^m$, where m was 1.38 for syphilis, 1.63 for HIV/AIDS, and 2 for measles or rubella, the frequency distribution of $N^{1.38}$, $N^{1.63}$, and N^2 , where N was population of municipalities, was compared with the frequency distributions of LIC sizes of syphilis, HIV/AIDS, measles, and rubella. The frequency distribution of LICs, particularly those of measles and rubella during the years when the epidemic was more severe, was close to the frequency distribution of N^m . The analysis suggested that LICs were products of stochastic events under the influence of municipality population size.

INTRODUCTION

How infectious diseases spread is the main theme of infection epidemiology. Though the Susceptible-Infected-Recovered model has been proposed (1), its application to epidemics in the real world is limited. For example, while there were 1,047 suspected cases of measles in Japan in 2014 (including 421 laboratory-confirmed cases and 93 cases linked to abroad travel [http://www.nih.go.jp/niid/images/iasr/rapid/meas/genotype/mv2014_20150917.pdf] Japanese), the epidemiological link was established only for 13 cases (2).

Since 1999, the National Epidemiological Surveillance of Infectious Diseases (NESID) has monitored the weekly incidence of infectious diseases in Japan under the Infectious Diseases Control Law (3). Such data can be presented in a table consisting of 47 rows showing the number of patients/week per prefecture and 52 columns representing the weeks of a year. This table is named the “Prefecture-Week Lattice” (“PW lattice”; an example is shown in Fig. 4C). We then defined “local infection cluster (LIC)” for measles and rubella as a cluster of infections reported from the same prefecture over the course of successive weeks, bordered by at least 1 week before and after, during which zero cases of infection were reported. We found that a majority of the clusters consisting of 1 patient, followed in order by

those consisting of 2 patients, 3 patients, and 4 or more patients. The size distribution of LIC was scale-free for measles and rubella (4).

In this report, we tested applicability of the LICs to other infectious diseases. Infections selected for the analysis were those for which all diagnosed cases are required to be reported under the Infectious Diseases Law of Japan (3); they included tetanus, severe invasive streptococcus infection with toxic shock syndrome (STSS), scrub typhus, legionellosis, measles, rubella, syphilis, and HIV/AIDS. Infections with annual frequency of fewer than 50 patients, or those acquired mostly abroad, e.g., typhoid fever (81% for 2005–2008) (5), shigellosis (69% for 1999–2000) (6), hepatitis A (80% for 2010–2014) (7), and giardiasis (43% for 2006–2013) (8), were excluded. Amebiasis and *enterohemorrhagic Escherichia coli* infection were not analyzed in this study, because the NESID data do not distinguish between cases acquired through person-to-person transmission and cases acquired through consumption of contaminated foods.

MATERIALS AND METHODS

Frequency distribution of LICs: The PW lattice (see Introduction) was used for the analysis. The size of LIC was determined by 2 parameters, “length” and “mass.” “Length” is the duration in weeks of a cluster and “mass” is the total number of patients per cluster (see reference 4 for detail). Thus, if the number of patients reported weekly by a prefecture appeared in the order of 0-0-2-3-1-0-..., the underlined group of figures “2-3-1” constitutes a cluster. Thus, cluster “2-3-1” had a mass of 6 and a length of 3. For “mass” and “length,” clusters were grouped according to size

Received January 4, 2016. Accepted April 6, 2016.
J-STAGE Advance Publication May 9, 2016.

DOI: 10.7883/yoken.JJID.2015.670

*Corresponding author: Mailing address: Department of Food Safety, Ministry of Health Labour and Welfare, 1-2-2 Kasumigaseki, Chiyoda-ku, Tokyo 100-8916, Japan.
E-mail: yoshikura-hiroshi@mhlw.go.jp

ranges, 2^0 , 2^1 , $>2^1-2^2$, $>2^2-2^3$, ... and the upper limit of the size range was used to represent the “size” of each group. The “frequency” of the “size” was calculated by adding together the number of clusters within the size range. For example, a cluster of size 2^3 consists of clusters of size 5, 6, 7, and 8. If the frequencies of respective sizes are 1, 0, 2, 0, for example, the frequency of the cluster is calculated as $1 + 0 + 2 + 0 = 3$.

In this study, we introduced a new parameter, “density,” which was obtained in principle by dividing mass by length. The frequency distribution of LICs was obtained by plotting the size of LICs in log2 scale on the x-axis and their frequency on the y-axis. The Zipf type plot (9,10) of the LIC sizes was created by plotting the LIC size on the x-axis and their ranking number (starting from the largest LIC) on the y-axis.

Simulation of frequency distribution of LICs: The size distribution of LICs produced by random events was simulated by tossing coins at random onto a 47×52 lattice as described in the previous report (4), which will be called “unweighted coin tossing.” We introduced a new simulation that took into account the dependency of disease incidence on population size. It was obtained by random tossing with probabilities proportional to the actual disease incidence in prefectures. For each coin toss, which corresponded to 1 infection, we assigned 2 random variables. The first random variable was taken from the $[0, 1]$ interval, which determined the prefecture of the incidence. Here, the $[0, 1]$ interval had been divided into 47 segments, whose widths were proportional to the weighted probabilities. The second random variable was taken from 1, 2, ..., 52, which determined the week of the incidence. All the simulations were obtained by 100 rounds of the coin tossing. The simulation will be called “coin tossing with weighted probabilities.”

Weekly incidence plot: The weekly incidence plot shows the number of patients per week for each prefecture from week 1 to week 52, which appears as a row in the PW lattice (See Fig. 4C). The number of patients was plotted on the y-axis and the week was plotted on the x-axis. The LICs were represented by peaks.

Approximations: The LIC frequency distribution was approximated by the log approximation (as the x-axis was in log2 and the y-axis was in log too, the plot was eventually a log-log plot), that for the Zipf type plots by the power approximation, and that for the weekly incidence by the moving average at an interval 2. All the calculations were performed with Microsoft Excel 2010.

RESULTS

Incidence of infectious disease versus population size: Fig. 1 shows the relationship between the incidence (number of patients/10,000,000 population) and population size ($\times 10,000,000$) of prefectures. If the number of patients was proportional to the population size, the plots would be horizontally distributed. If the number of patients per prefecture was similar among prefectures, the plots would be distributed towards the upper left because the denominator decreases to the left side (called “even” type). If the spread of infection depended on larger population size, the plots would

be distributed towards the upper right (called “population-driven” type).

Tetanus and STSS were “even” type (Figs. 1A and 1B). Rubella, measles, HIV/AIDS, and syphilis were “population-driven” type (Figs. 1E, 1F, 1G, and 1H). The population dependency was stronger for rubella and measles than for syphilis and HIV/AIDS (slope was 1.37–1.86 for the former and 0.11–0.58 for the latter). The plots of scrub typhus and legionellosis deviated neither to the left nor to the right (Figs. 1C and 1D).

LIC size distribution of tetanus, severe invasive STSS, scrub typhus, and legionellosis: These infections are not transmitted from person-to-person except for STSS in rare occasions.

Tetanus was acquired by wound injury and its occurrence was essentially sporadic (see Fig. 2-A3: plots for Hokkaido and Kagoshima that reported relatively large numbers of patients in 2013 and 2000, respectively). The slope of the LIC frequency distribution was -2.19 – -2.37 (Fig. 2-A1), which was close to -2.58 – -2.44 , the slope expected from the coin tossing with unbiased probabilities (4).

STSS is an acutely progressing septic shock caused by β -hemolytic streptococci. The infection occurs among patients with pharyngitis or a skin wound. It is generally sporadic, but may occur as clusters, for example, in care facilities (11). The initial slope was steep, -1.75 – -1.96 (Fig. 2-B1), which was similar to the pattern obtained by the coin tossing with unbiased probabilities. The flat tailing in Fig. 2-B1 was attributable to the clustered infections.

Scrub typhus was transmitted by the *Orientia tsutsugamushi* vector during farm work or other outdoor activities in the endemic regions. There was 1 epidemic season from week 45 to week 10 of the next year in Hiroshima, Gifu, Kagoshima, Miyazaki, and other areas in the Southern part of Japan, (Fig. 2-C3, upper panel), while there were 2 epidemic seasons in weeks 20–30 and 40–50 weeks in Aomori, Akita, Fukushima, and other areas in the Northern part of Japan (Fig. 2-C3, lower panel). The occurrence was essentially sporadic, but group activities may result in clustered occurrence. The patients were mostly elderly (70–74 years) (12). Size frequency distribution of LICs had an initial steep slope, -1.64 , which was compatible with the random event (Fig. 3-A1). The milder slope that followed the steep slope was attributable to the less frequent clustered infections.

Legionellosis is caused by inhalation of *Legionella* spp. It is not transmitted from person-to-person. In 2000, there were only 2 large peaks attributable to outbreaks in spas (Fig. 2-D3 upper panel), but in 2013 there were multiple peaks (Fig. 2-D3 lower panel). From 1979 until 1992, only 86 cases were reported, but since 2008, the annual incidence increased to 800–900 cases, including many outbreaks in spas, welfare facilities, hotels, gymnastic facilities, and other places using filtered and circulated water (13). The steep slope in 2000, -1.78 , (Fig. 2-D1), was compatible with the sporadic occurrence of this infection.

The slope of Zipf type plots was -2.41 – -4.33 for tetanus and STSS, while it was -0.95 – -1.70 for scrub typhus and legionellosis (Figs. 2-A2, 2-B2, 2-C2, and

Fig. 1. Relation between population size and number of patients for different prefectures. The horizontal axis is the population size of prefectures ($\times 10,000,000$), and the vertical axis number of patients/10,000,000 population per year for each prefecture. The approximations in the panels for rubella, measles, rubella, HIV/AIDS are power approximations obtained for the log-log plots in the Excel file. Total number of patients per year was 91 and 128 for tetanus (2000 and 2013), 47 and 203 for STSS (2000 and 2013), 2,386 and 14,344 for rubella (2012 and 2013), 11,012 and 732 for measles (2008 and 2009), 794 and 1,586 for HIV/AIDS (2000 and 2013), and 759 and 1,228 for syphilis (2000 and 2013).

2-D2). The milder slope for the latter is attributable to the more frequent occurrence in clusters. The plot pattern of these infections was essentially what is expected from the sporadic nature of these infections.

LIC size distribution of rubella and measles: Rubella and measles are highly infectious diseases, which are transmitted from person-to-person by droplets. Due to the short incubation and short communicability period (~ 2 weeks), the time interval between the infection and notification is several days. Both infections were of the population-driven type (Figs. 1E and 1F). The frequency distribution of the LICs (Figs. 3-A1 and 3-B) was reproduced from the previous report (4). The plots fell on the rightward descending straight line. The slope was -0.38 and -0.74 for measles in 2008 and 2009, and -0.48 and -0.52 for rubella in 2012 and 2013, respectively. The slope of the Zipf type plot was -0.58 and -1.21 for measles in 2008 and 2009, and -0.66 and -0.54 for rubella in 2012 and 2013 (4). The year of a severe epidemic for measles was 2008 and for rubella the year was 2013. The slope was milder for these epidemic years both for measles and rubella.

Fig. 3-A2 shows the weekly incidence of measles in prefectures that reported large numbers of patients. In

2008, a sharp large peak of Fukuoka appeared at week 5 (panel M08a), followed by broader peaks in Kanagawa in weeks 7–11, in Tokyo in week 9 (panel M08b), in Osaka in weeks 7 and 18 (panel M08c), in Hiroshima in week 11, in Okayama in week 20 (panel M08a), in Shizuoka in week 12, in Aichi in weeks 12 and 20 (panel M08d), and in Kyoto in week 23 (panel M08c). The pattern accurately reflected the propagation of the measles epidemic from one area to another. In year 2013, there was a small peak of Hiroshima in weeks 4–6 (panel M13b), followed by the large peak of Tokyo in weeks 14–24 (panel M13a). No other prefectures experienced large outbreaks, which may be due to the success of the measles elimination program that was intensified in 2008 (14).

Fig. 3-B2 shows the weekly incidence of rubella in 2012 and 2013. In 2012, the first peak was that of Tokyo in week 26 (panel R12a), which was followed by peaks in Osaka and Hyogo in weeks 29–33 (panel R12b), and that of Fukuoka in week 37 (panel R12c). In 2013, the epidemic season started in Tokyo as a broad peak (the peak was in week 15); the epidemic then shifted to Kanagawa, Saitama, and Chiba, nearby prefectures of Tokyo (panel R13a). In week 22, a broad

Fig. 2. LIC size frequency analysis of tetanus, STSS, scrub typhus, and legionellosis. A: tetanus; B: STSS; C: scrub typhus; D: legionellosis. The number that follows the large alphabet indicates: 1: Frequency distribution of LIC mass (LIC mass in the x-axis and the frequency in the y-axis with the approximations for the initial slopes); 2: the Zipf type plot (LIC mass in the x-axis and the ranking number in the y-axis with approximations); and 3: weekly incidence (weeks in the x-axis and incidence/week in the y-axis). The equations under the graphs A1, B1, C1, and D1, $y = ke^{sx}$, indicate exponential approximation of relation between cluster size (x) and its frequency (y). The equations under graphs A2, B2, C2, and D2, $y = kx^{-s}$, indicate power approximation of relation between cluster size of LICs (x) and their ranking number (y).

peak appeared in Osaka and Hyogo (panel R13b), and in week 24 in Fukuoka and Aichi (panel R13c), distant from Tokyo.

The plot of the weekly incidence, from which LICs were derived, accurately replicated the geographic spread of measles and rubella.

LIC size distribution of syphilis and HIV/AIDS: Syphilis and HIV/AIDS are acquired primarily through sexual contact. They were identified as population-driven types (Figs. 1G and 1H). In contrast to measles and rubella, that are reported within a short time interval after infection, they may be reported at later stages of infection. Nevertheless, the majority of them were reported in the earlier phase of infection. In 2013, for example, among a total 1,228 syphilis cases reported, 692 were early symptomatic, and 466 were asymptomatic; only 66 were late symptomatic and the remaining 4 were congenital syphilis (15). Among 1,590 HIV/AIDS cases, 1,106 were detected before development of AIDS, while 484 were detected after the development of AIDS (16).

As the weekly incidence in the NESID data did not distinguish the early from the late cases, the data were a mix of both. Therefore, HIV/AIDS and syphilis are expected to have LIC size distributions that are different from those of measles and rubella.

Figs. 4A and 4B show the size distribution of LICs of syphilis and HIV/AIDS, respectively. The slopes were -0.72 – -0.62 for HIV/AIDS and -0.75 – -0.70 for syphilis (Figs. 4-A1 and 4-A2). The slopes of syphilis and HIV/AIDS were almost identical to those of measles in 2009 (-0.74) and rubella in 2012 (-0.85). The slopes of the Zipf type plot was -1.74 – -1.16 for HIV/AIDS and -1.43 – -1.28 for syphilis (Figs. 4-B1 and B2).

Figs. 4-A3 and 4-B3 show the weekly incidence plots of syphilis and HIV/AIDS for 2 prefectures, with the highest frequencies in 2013. The epidemic continued throughout the year, and the plot pattern was uncharacteristic, with randomly distributed peaks of similar sizes. It was surprising that the frequency distribution pattern of LICs of HIV/AIDS and syphilis resembled

Fig. 3. LIC size frequency analysis of measles and rubella. A: measles in 2008 and 2009; B: rubella in 2012 and 2013. A1 and B1: Frequency distribution of LIC mass for measles and rubella, respectively (LIC mass in the x-axis and the frequency in the y-axis). A2 and B2: weekly incidence plots for measles and rubella, respectively (weeks in the x-axis and incidence/week in the y-axis).

that of measles and rubella, because their transmission routes and the weekly incidence pattern were substantially different.

Fig. 4C shows the PW-plot of HIV/AIDS in 2013, which can be compared to that of measles shown in the previous publication (4).

Simulation of the scale-free distribution by random coin tossing with weighted probabilities: The above observations led us to hypothesize the random nature of HIV and syphilis infections. We simulated a situation where 5 prefectures received coins at a higher probability than the remaining 42 prefectures, and 125–1,000 coins were tossed at random. As shown in Fig. 5, the frequency distribution of LICs fell on straight lines under conditions where the infection probability in the 5 prefectures was 2.5- or 5-fold higher than the rest (coded as $P_H = 2.5 \times P_L$ and as $P_H = 5 \times P_L$, respectively) (Fig. 5). The slope for $P_H = 5 \times P_L$, -1.95 – -0.83 , partly overlapped the range of the slopes of HIV/AIDS, syphilis, measles, and rubella, which was -0.85 – -0.38 .

The data thus far presented are summarized as a

table in Fig. 6A. In Fig. 6B, the slopes of the frequency distribution were plotted against the number of patients. The slope became milder as the total number of patients increased, for both actual and simulated epidemics, though the slope of the simulation appeared steeper. The simulation was further refined by using probabilities adjusted to the real epidemics. The 47 prefectures were divided into 37 prefectures with relative probability $\times 1$, 5 prefectures with relative probability $\times 4$, and the remaining 5 prefectures with relative probability $\times 16$ (Fig. 7B), where the probabilities were made approximately proportional to the actual number of patients per prefecture (refer to Fig. 7A). Fig. 7C shows the frequency distribution of LICs thus obtained. Trials with 800 or 2,000 tosses, respectively, produced straight lines with slopes -0.65 and -0.82 , which were close to the slope of measles in 2009, -0.74 (Fig. 7C and 7D), and the slopes of HIV/AIDS (-0.62 – -0.72), and syphilis (-0.70 – -0.75) (Fig. 6A).

The weekly incidence plots of the epidemic simulated with 800 tosses (Fig. 7E) was uncharacteristic, with the peaks of similar sizes appearing randomly; the pattern

Fig. 4. LIC size frequency analysis of HIV/AIDS and syphilis. A: HIV/AIDS in 2000 and 2013; B: syphilis in 2000 and 2013. A1 and B1: Frequency distribution of LIC mass for HIV/AIDS and syphilis (LIC mass in the x-axis and the frequency in the y-axis); A2 and B2: the Zipf type plots for HIV/AIDS and syphilis (LIC mass in the x-axis and the ranking number in the y-axis). The equations under the graphs A1 and B1, $y = ke^{sx}$, indicate exponential approximation of relation between cluster size (x) and its frequency (y). The equations under graphs A2 and B2, $y = kx^{-s}$, indicate power approximation between cluster size of LICs (x) and their ranking number (y). A3 and B3: weekly incidences for HIV/AIDS and syphilis (weeks in the x-axis and incidence/week in the y-axis); C: PW plot of HIV/AIDS in 2013, which could be compared with the PW plot of measles (4).

resembled the weekly incidence of syphilis and HIV/AIDS.

It was interesting to note that, in the simulation of LICs, when the number of tosses was increased to 5,000 or 10,000, the plots became wavy (Fig. 7C). In the actual epidemics, however, the frequency distributions were on straight lines (see the plot of measles in 2008 and the plot of rubella in 2013, Fig. 7C). This discrepancy was probably brought about by the strong influence of the municipality population size on the disease incidences (see below).

Length-density plot: The LIC size distribution of syphilis-HIV/AIDS (Figs. 4-A1 and 4-B1), and that of measles-rubella (Fig. 3-A1 and Fig. 3-B1), were almost identical in spite of their different transmission routes and different epidemiological characteristics (compare weekly plots in Figs. 3-A2 and 3-B2 with those in Figs. 4-A3 and 4-B3). In order to find differences in LICs of measles and rubella compared to syphilis and HIV/AIDS, a new parameter, “density,” was introduced.

The “density” was defined as “mass” divided by

“length.” For analysis, we prepared 3 tables. First, frequencies of the length-mass pairs were tabulated as shown in Fig. 8-A1, where iN_j denotes frequency of an LIC with length i and mass j . Then, $iN_j \times j$, the total mass for each length-mass pair, was tabulated as shown in Fig. 8-A2. The total length $\sum iN$ and the total mass M_i for length i were calculated, respectively, by adding together total iN_j 's and total $iN_j \times j$'s (2nd and 3rd columns of Fig. 8-A3, respectively). Basic mass, B_i , was the sum total of the lengths of all the LICs for length i (4th column of Fig. 8-A3). Density was obtained by dividing the M_i by the B_i (5th column of Fig. 8-A3). The length was plotted on the horizontal axis and the density on the y-axis (Fig. 8-B1 for syphilis and HIV/AIDS and Fig. 8-B2 measles and rubella; the plots for the random tosses of 800, 2,000, or 10,000 coins were inserted in both graphs). In the case of syphilis and HIV/AIDS, the plots of the actual epidemics and the simulated epidemics were intermingled. For measles and rubella, the plots of the actual epidemics were distributed far above the simulated epidemics,

Scale Free Distribution of Infection Cluster Sizes

Fig. 5. Simulation with random coin tosses on the 47×52 lattice. Five prefectures among 47 were allocated with probability of receiving coins higher than the remaining 42 prefectures by 1.25 ($P_H = 1.25 \times P_L$), by 2.5 ($P_H = 2.5 \times P_L$), by 5 ($P_H = 5 \times P_L$), or by 10-fold ($P_H = 10 \times P_L$). The number of coins tossed was 125 ($\square$), 250 ($\triangle$), 500 ($\circ$), or 1,000 ($\diamond$). The plot was the result of 100 rounds of the tosses. The mass of LICs is plotted in the x-axis in the $\log_2$ scale and the frequency of LIC sizes in the y-axis.

i.e., the density of the actual epidemics was higher than that of the simulated epidemics for measles and rubella, while the density of the actual epidemics and that of the simulated ones were similar for syphilis and HIV/AIDS.

The observation that the density of LICs of measles and rubella was higher than that of syphilis and HIV/AIDS was consistent with the epidemiological properties of these epidemics, i.e., the epidemic of measles and rubella was concentrated in time and space, while that of syphilis and HIV/AIDS was diffusely dispersed in time. The observation that the epidemics of syphilis and HIV/AIDS resembled the simulated epidemics in the distribution of size and mass of LICs, and in the weekly plots, indicated that the epidemics of syphilis and HIV/AIDS were essentially stochastic under the weighted probabilities.

Deriving scale free distribution of LICs from frequency distribution of population size of local communities: As the weighted probabilities were derived from the observation that the number of patients was higher in more populated prefectures, and, as we suspected, the LICs reflected virus spread in the municipality level rather than in the prefecture level, we tried to simulate the size distribution of LICs in the actual epidemics based on the size distribution of the municipality populations.

The cumulative frequency distribution of the popula-

tion sizes of 1,741 municipalities in Japan (cities, towns, and villages, which are the administrative units in Japan) is shown in Fig. 9-A1, and that of population sizes of municipalities in Tokyo, Chiba, Nagoya, Osaka, and Fukuoka in Fig. 9-A2. The distribution was sigmoid and compatible with the lognormal distribution (the initial long mild slope for Tokyo was attributable to several lowly populated island municipalities). Thus, the lognormal distribution of population sizes of the municipalities in Japan was replicated in the lognormal distribution of population sizes of municipalities in the individual prefectures. Previous data show lognormal distribution of population size of prefectures (17). Therefore, the population distribution pattern was repeated at every scale of Japanese communities, suggesting that population size had characteristics of fractals.

For 238 municipalities with population size larger than 120,000 (encircled by larger symbols), the plots fell on straight lines, with equations $y = 9215x^{-1.42}$ for Japan (Fig. 9-B1), and $y = 59x^{-1.38}$ and $y = 25x^{-1.36}$ for Tokyo and Chiba (Fig. 9-B2), respectively. Namely, the size distribution of these large municipalities was scale-free. Though the plots of the large municipalities fell on straight lines for at least 2 orders of magnitude, which is a rule of thumb of candidate power law, if smaller municipalities were included, the distribution becomes lognormal. Such an observation has already

Fig. 6. Summary of number of patients/ year, slope of the LIC frequency distribution for various infections and simulations. A: Summary table. “Simulation” in the bottom of the table indicates simulation for infections asterisked under the simulation conditions tabulated in Fig. 7B. B: Graphic presentation of the table in Fig. 6-A. Total number of patients in the logarithmic scale is plotted in the x-axis and the slopes of the frequency distribution of LICs (in the normal scale) in the y-axis.

been made by Eeckhout regarding the 2,000 United States Census data (18).

In Fig. 9-C1, the population sizes of the 217 large municipalities (population $\geq 20,000$) are plotted on the x-axis (in $\log_2$) and their frequencies on the y-axis. The plots fell on a straight line with equations $y = 3,643e^{-0.91x}$ (○). Fig. 9-C2 shows similar plots for Tokyo ($y = 164e^{-0.58x}$) and for Chiba ($y = 211e^{-0.86x}$).

The number of patients (P) and population size (N) could be correlated with an equation $P = kN^m$, where m was ~ 2 for measles and rubella (17), 1.63 for HIV/AIDS and 1.38 for syphilis (19). We suspected that the frequency distribution of these infections could be derived from the frequency distribution of population size (N) by replacing N (population size) with N^2 and $N^{1.63}$ and $N^{1.38}$, respectively. As shown in Fig. 9-C1, the slope of the size distribution of the population size N (○) was -0.91 ; and those of N^2 (●), $N^{1.63}$ (▲), and $N^{1.38}$ (■) were -0.44 , -0.58 , and -0.64 , respectively. The slope of N^2 , -0.44 , was close to that of measles in 2008, which was -0.38 (Fig. 3-A1), and that of rubella in 2013, which was -0.52 (Fig. 3-B1). The slope of $N^{1.63}$, which was -0.58 , was also close to the slope of HIV/AIDS, which was -0.62 – -0.72 (Fig.

4-A1), and the slope of $N^{1.38}$, which was -0.64 , was near that of syphilis, -0.70 – -0.75 (Fig. 4-B1).

In other words, the size distribution of LICs was very similar to the size distribution of N^m that was derived from the size distribution of the municipality population sizes. It may imply that LICs represent the municipality level epidemics. This possibility could have been checked by the municipality level data of disease incidence, but such data are not available from the current NESID database.

It is of interest to note that, for measles and rubella, while the epidemics with large numbers of patients (slopes -0.38 for measles in 2008, and -0.52 for rubella 2013) were nicely simulated by the size distribution of municipality sizes whose slope was -0.49 , the epidemics with smaller patient numbers in other years, whose slopes were -0.52 – -0.89 for measles in 2010–2013 and -0.89 – -1.5 for rubella in 2008–2011, could be better simulated by the 800 random coin tossing with the weighted probabilities, whose slope was -0.68 . In other words, while the frequency distribution of LICs of large epidemics was simulated better by using the frequency distribution of the municipality population size (compare plots in Fig. 9-C1 with wavy

A: Cumulative frequency distribution

D: Comparison of actual frequency distribution of LICs of measles with simulated plots

B: Simulation conditions

Total tosses (patients)	No. of prefectures	Relative probability
800	37	1
	5	4
	5	16

C: Frequency distribution of LICs in the Simulated population

E: Weekly incidence plot of the simulated epidemic

Fig. 7. Random tossing under the condition of the weighted probabilities. A: cumulative frequency distribution of number of patients per prefecture for rubella in 2013, measles in 2008, syphilis in 2000, and HIV in 2000. The population was divided into 3 groups: relative probability $\times 16$ was allocated to 5 prefectures (indicated P: $\times 16$ on the left side of the graph), relative probability $\times 4$ to 5 prefectures with (about 5 prefectures with the range indicated P: $\times 4$ on the left side of the graph) and relative probability $\times 1$ to 37 prefectures (range indicated P: $\times 1$ on the left side of the graph). The range of number of patients can be read by extrapolating the downward arrows to the x-axis. B: conditions used for the coin tossing. In this simulation, relative probability $\times 4$, for example, was produced by using boxes 4-fold larger than boxes for probability $\times 1$. C: frequency distribution of LICs obtained after 100 rounds of coin tossing under the condition given in Fig. 7B. $\blacksquare$: 800 tosses per round (patients); $\bullet$: 2,000 tosses per round (patients); $\triangle$: 5,000 tosses per round (patients); and $\diamond$: 10,000 tosses per round (patients). D: comparison of the frequency distribution of the actual measles epidemics with the simulated ones with 800 tosses per round. E: weekly incidence plots for the simulated epidemics; the plots for epidemics in simulated prefectures with different number of patients are shown. “Simulation row #” refers to a prefecture in the simulated epidemics (the total number of patients/prefectures in the parenthesis).

plots for 5,000 and 10,000 coin tosses in Fig. 7C), the smaller epidemics were better simulated by the random tossing under weighted probabilities (Fig. 7D). This appears to indicate that, during more severe epidemics, the population size of municipalities was the major determinant of the size distribution of epidemics, while, during smaller epidemics, random events with weighted probability played the major role.

DISCUSSION

The size distribution of LICs of infections that are transmitted from person-to-person, such as measles, rubella, HIV/AIDS, and syphilis, fell on straight lines of at least 2 orders of magnitude, a rule of thumb of

candidate power law (20). This size distribution was entirely different from that of LICs of infections that were not transmitted from person-to-person, such as tetanus, STSS, scrub typhus, and legionellosis. The frequency distribution of the former was scale-free, while that of non-person-to-person transmission was what is expected from geographically random distribution. The weekly incidence plots for infections with person-to-person transmission was consistent with the hypothesis that the LICs represented the local spread of infection. The inference was consistent with the common knowledge that infections spread most easily to the nearest people in the same communities.

The scale-free distribution of LIC sizes of measles, rubella, HIV/AIDS, and syphilis could be simulated

A1

Cluster length (L)	No. patients per cluster for each length-mass pair						
1	1N1	1N2	1N3	1N4	1N5		
2	x	2N2	2N3	2N4	2N5		
3	x	x	3N3	3N4	3N5	3N6 x 6	
4	x	x	x	4N4	4N5	4N6 x 6	

A2

Cluster length (L)	Mass (patients/cluster) x frequency for each length-mass pair						
1	1N1 x 1	1N2 x 2	1N3 x 3	1N4 x 4	1N5 x 5		
2	x	2N2 x 2	2N3 x 3	2N4 x 4	2N5 x 5		
3	x	x	3N3 x 3	3N4 x 4	3N5 x 5	3N6 x 6	
4	x	x	x	4N4 x 4	4N5 x 5	4N6 x 6	

A3

Cluster length (L)	Total length for respective length (ΣL)	Total mass for respective length (Mi)	Basic mass (Bi)	Density (Mi/Bi)
1	1N1 + 1N2 + 1N3 + ... = $\Sigma 1N$	1N1x1 + 1N2x2 + 1N3x3 + ... = M1	1 x $\Sigma 1N$ = B1	M1/B1
2	2N2 + 2N3 + 2N4 + ... = $\Sigma 2N$	2N2x2 + 2N3x3 + 2N4x4 + ... = M2	2 x $\Sigma 2N$ = B2	M2/B2
3	3N3 + 3N4 + 3N5 + ... = $\Sigma 3N$	3N3 x3 + 3N4x4 + 3N5x5 + ... = M3	3 x $\Sigma 3N$ = B3	M3/B3
4	4N4 + 4N5 + 4N6 + ... = $\Sigma 4N$	4N4x4 + 4N5x5 + 4N6x6 + ... = M4	4 x $\Sigma 4N$ = B4	M4/B4

B1

B2

Fig. 8. Derivation of density and plot of density vs. length. A: derivation of the density. See text for detail. B: Plot of “density” vs. “length” for epidemics of syphilis and HIV/AIDS (B1) and for rubella and measles (B2). The plots for real epidemics are shown with closed symbols and those of the simulated epidemics with open symbols. The length is plotted in the x-axis and the density in the y-axis both in the logarithmic scale.

by:

- postulating random events with weighted probabilities, which were higher for populated prefectures; and
- Using an equation $P = kN^m$, where P and N stood for number of patients and population size respectively, and m for ~ 2 for measles and rubella, 1.63 for HIV/AIDS, and 1.38 for syphilis.

For measles, in 2009, when the number of patients was 447, the slope was -0.74 , while in 2008, when the number of patients was 11,012, the slope was -0.38 . For rubella, in 2012, when the number of patients was 2,386, the slope was -0.48 , while in 2013, when the number of patients was 14,433, the slope was -0.52 (4). The simulation postulating random events with slope -0.65 or -0.82 (Fig. 7D) fit well to the years with a less severe epidemic (slopes -0.74 and -0.85 for measles and rubella respectively), and the simulation using the equation $P = kN^2$ with slope -0.44 fit better to the years with a more severe epidemic (slopes -0.38 and -0.48 for measles and rubella, respectively). It is likely that when the number of patients is small, the infection events are largely at random, but when the epidemic becomes widespread, affecting the entire community, the population size of local communities may dominantly determine the frequency distribution of LIC sizes.

The number of patients with HIV/AIDS was 891 in 2000 and 1,590 in 2013 and the slopes for those years were -0.72 and -0.62 , respectively. The number of patients with syphilis was 769 in 2000 and 1,586 in 2013, and the slopes for those years were -0.75 and -0.70 , respectively. These epidemics could be simulat-

ed fairly well by postulating random events (simulated slope -0.65 or -0.82 in Fig. 7D), as well as by using formula $P = kN^{1.63}$ (simulated slope -0.58 vs. the actual slope -0.72 or -0.62) for HIV/AIDS, and the formula $P = kN^{1.38}$ (simulated slope -0.64 vs. actual slope -0.70 or -0.75) for syphilis. Thus, *grosso modo*, the random coin tossing with the weighted probabilities, and the simulation using the size distribution of municipality populations, gave nearly the same distribution patterns and accurately replicated the frequency distribution of LICs. As the weighted probability was larger for populated communities, it was indicated that the size distribution of LICs was determined essentially by the size distribution of human communities, which was lognormal (Fig. 9A). The lognormal distribution is the multiplicative product of many independent random variables, each of which is positive. It is observed that the cities grow in proportion to their sizes (21).

Communities with a larger population tend to increase in size, while those with a smaller number of people remain stable in size. The pathogens carried by humans are more attracted to populated municipalities, where there is a large population available to infect.

Conflict of interest None to declare.

REFERENCES

- Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford University Press; 1991. p. 17-9.
- World Health Organization Western Pacific Region. Regional and country profiles of measles elimination. Data through December 2015 (as of May 2016).

Scale Free Distribution of Infection Cluster Sizes

Fig. 9. Population size distribution of municipalities in Japan as of 1 October 2015 (http://uub.jp/rnk/cktv_j.html). A1: cumulative frequency distribution of population size of all the municipalities in Japan; A2: cumulative frequency distribution of population size of the municipalities in Tokyo, Chiba, Nagoya, Osaka, and Fukuoka. B1: the Zipf type plot of population size of all the municipalities in Japan; the plots for 238 municipalities with population > 12,000 are enclosed in larger symbols; the straight line is the power approximation for municipalities with population > 12,000. B2: Zipf plot of population size of the municipalities in Tokyo and Chiba (the approximations are for municipalities with population > 12,000). C1: frequency distribution of population size ($\times 10,000$) of municipalities for municipalities > 12,000 (○), that of $N^{1.4}$ (▲) and that of N^2 (●). C2: frequency distribution of population size ($\times 10,000$) at $\log_2$ intervals of municipalities for municipalities > 12,000 in Tokyo (○) and Chiba (△). The x-axis indicates population size in $\log_2$ and the y-axis the frequency.

- Ministry of Health Labour and Welfare. Notification of patients under the Infectious Disease Control Law. Available at http://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryuu/kenkou/kekaku-kansenshou/kekaku-kansenshou11/01.html. Retrieved in December 2015. Japanese.
- Yoshikura H, Takeuchi F. Measles and rubella-scale free distribution of local infection clusters. *Jpn J Infect Dis.* 2016;69:293-9.
- National Institute of Infectious Diseases and Tuberculosis and Infectious Diseases Control Division, Ministry of Health, Labour and Welfare. Typhoid fever and paratyphus 2005-2008. *Infect Agents Surveillance Rep.* 2009;30:91'-2'.
- National Institute of Infectious Diseases and Tuberculosis and Infectious Diseases Control Division, Ministry of Health, Labour and Welfare. Shigellosis 1999-2000. *Infect Agents Surveillance Rep.* 2001;22:81'-2'.
- National Institute of Infectious Diseases and Tuberculosis and Infectious Diseases Control Division, Ministry of Health, Labour and Welfare. Hepatitis A in Japan, 2010-2014, as of November 2014. *Infect Agents Surveillance Rep.* 2015;36:1'-2'.
- National Institute of Infectious Diseases and Tuberculosis and Infectious Diseases Control Division, Ministry of Health, Labour and Welfare. Cryptosporidium and Giardiasis as of July 2014. *Infect Agents Surveillance Rep.* 2014;35:185'-6'.
- Newman ME. Power laws, Pareto distribution and Zipf's law. *Contemporary Phys.* 2005;46:323-51.
- Newitz A. A mysterious law that predicts the size of the world's biggest cities Available at <http://io9.com/the-mysterious-law-that-governs-the-size-of-your-city-1479244159>. Retrieved in December 2015.
- National Institute of Infectious Diseases and Tuberculosis and Infectious Diseases Control Division, Ministry of Health, Labour and Welfare. Streptococcal infections in Japan, April 2006-2011. *Infect Agents Surveillance Rep.* 2012;33:209'-10'.
- National Institute of Infectious Diseases and Tuberculosis and Infectious Diseases Control Division, Ministry of Health, Labour and Welfare. Scrub typhus and Japanese spotted fever in Japan 2006-2009. *Infect Agents Surveillance Rep.* 2010;31:120'-1'.
- National Institute of Infectious Diseases and Ministry of Health Labour and Welfare, Tuberculosis and Infectious Disease Con-

- trol Division: Legionellosis, January 2008–December 2012, Infect Agents Surveillance Rep. 2013;34:155'-6'.
14. National Institute of Infectious Diseases and Ministry of Health Labour and Welfare, Tuberculosis and Infectious Disease Control Division. Measles in Japan, as of March 2015. Infect Agents Surveillance Rep. 2015;36:51'-2'.
 15. National Institute of Infectious Diseases and Ministry of Health Labour and Welfare, Tuberculosis and Infectious Disease Control Division. Syphilis 2008–2014. Infect Agents Surveillance Rep. 2015;36:17'-8'.
 16. National Institute of Infectious Diseases and Ministry of Health Labour and Welfare, Tuberculosis and Infectious Disease Control Division. HIV/AIDS in Japan, 2013. Infect Agents Surveillance Rep. 2014;35:203'-4'.
 17. Yoshikura H. Impact of population size on incidence of rubella and measles in comparison with that of other infectious diseases. Jpn J Infect Dis. 2014;67:447-57.
 18. Eeckhout J. Gibrat's law for (all) cities. The Amer Econ Rev. 2004;94:1429-51.
 19. Yoshikura H. Geo-demography of HIV/AIDS in Japan from 1985 to 2011–Incidence and transmission mode under influence of population size/density. Jpn J Infect Dis. 2016;69:97-108.
 20. Stumpf MP, Porter MA. Mathematics. Critical truths about power laws. Science 2012;335:665-6.
 21. Batty M. The New Science of Cities. Cambridge, MA, The MIT Press; 2013. p. 151-77.