Communications for Statistical Applications and Methods
2014, Vol. 21, No. 6, 487–500

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.6.487
Print ISSN 2287-7843 / Online ISSN 2383-4757

Estimating the Transmittable Prevalence of Infectious Diseases Using a Back-Calculation Approach

Youngsaeng Lee[a], Hyun Gap Jang[b], Tae Yoon Kim[c], Jeong-Soo Park[1,d]

[a]ICT Convergence Research Center for Smart Appliances, Chonnam National University, Korea
Current at: Dept. of Math. & Stat., University of British Columbia, Kelowna, Canada
[b]JW LEE Center for Global Medicine, College of Medicine, Seoul National University, Korea
[c]Department of Statistics, Kyemyong University, Korea
[d]Department of Statistics, Chonnam National University, Korea

Abstract

A new method to calculate the transmittable prevalence of an epidemic disease is proposed based on a back-calculation formula. We calculated the probabilities of reactivation and of parasitemia as well as transmittable prevalence (the number of persons with parasitemia in the incubation period) of malaria in South Korea using incidence of 12 years(2001–2012). For this computation, a new probability function of transmittable condition is obtained. The probability of reactivation is estimated by the least squares method for the back-calculated long-term incubation period. The probability of parasitemia is calculated by a convolution of the survival function of the short-term incubation function and the probability of reactivation. Transmittable prevalence is computed by a convolution of the infected numbers and the probabilities of transmission. Confidence intervals are calculated using the parametric bootstrap method. The method proposed is applicable to other epidemic diseases in other countries where incidence and a long incubation period are available.

We found the estimated transmittable prevalence in South Korea was concentrated in the summer with 276 cases on a peak at the 31st week and with about a 60% reduction in the peak from the naive prevalence. The statistics of transmittable prevalence can be used for malaria prevention programs and to select blood transfusion donors.

Keywords: Epidemiologic methods, incubation period, malaria, least squares method, parasitemia, survival function, transmission.

1. Introduction

Malaria remains a problematic disease of the 21st century (WHO, 2005). It kills more than one million people each year in the world. Global climatic change will allow malaria to spread into northern latitudes such as Europe and large parts of the United States (Rogers and Randolph, 2000). It is caused by a protozoan parasite in the phylum, Apicomplexa, and in the genus, Plasmodium. There are four species that are in the genus: Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale, and Plasmodium malariae (Greenwood and Mutabingwa, 2002). Two species of these, P. vivax and P. ovale, tend to have a hypnozoites stage and long incubation period (CDC, 2006).

P. vivax in South Korea was highly endemic until 1910 and decreased gradually after the application of modern medicine. It was thought to be eradicated after 1984. But malaria reemerged in

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A4A01009355).

[1] Corresponding author: Department of Statistics, Chonnam National University, Gwangju 500-757, Korea.
E-mail: jspark@jnu.ac.kr

Published 30 November 2014 / journal homepage: http://csam.or.kr

the demilitarized zone between North and South Korea, after 1993 because of the shortage of malaria eradication programs in North Korea (Park *et al.*, 2005; Lee *et al.*, 2002). Around 2,000 people are infected annually (KCDC, 2007).

We assume that if some diseases with a seasonal fluctuation have a long incubation period, their infection curve would be different from the incidence curve. Malaria in South Korea satisfies those two requirements. P. vivax (the only species in Korea) has a long incubation period and clear seasonality that reflects the population dynamics and other entolological characteristics of the Anopheles sinensis vector that hibernates during the winter season (Burket *et al.*, 2002; Ree *et al.*, 2001).

In this study, we computed the probabilities of reactivation and parasitemia, and the transmittable prevalence (the number of persons with parasitemia in the incubation period). For this computation, a new method based on the back-calculation formula is proposed. The probability of reactivation is estimated by the least squares method for the back-calculated long-term incubation period. The probability of parasitemia is calculated by convolution of the survival function of the short-term incubation function and the probability of reactivation. Transmittable prevalence is computed by a convolution of the infected numbers and the probability of transmission. Details of these computations are provided in Sections 3 and 4.

The back-calculation method, a major technique described in this paper, has been used to calculate annual HIV infections from the annual incidence, their incubation distribution and other information (Brookmeyer and Gail, 1988; Bacchetti *et al.*, 1993; Hall *et al.*, 2008; Punyacharoesin and Viwatwongkasem, 2009). The method has also been used to estimate the number of dependent heroin users in Australia (Law *et al.*, 2001) and long-term trends in the incidence and prevalence of opiate use/injecting drug use in England for 1968–2000 (De Angelis *et al.*, 2004). It was used to estimate the number of SARS cases imported by international air travel (Goubar *et al.*, 2009), and age specific cancer incidence rates (Mezzetti and Robertson, 1999). In this study, details to estimate infection distribution using back-calculation formula are described in the Appendix because the formula is already known.

2. Back-calculation and Prevalence of Malaria

2.1. Data

All medical facilities in South Korea should report their malaria cases to public health centers and then to the Korean Center for Disease Control(KCDC). We used their reporting data from 2001 to 2012 for our incidence data (KCDC, 2008) because the KCDC service tracked daily incidence days after mid 2000. Figure 1 shows the time series of reported cases for 12 years. We used only domestic malaria infection and excluded all overseas infection. A total of 17,280 cases were reported entirely for 12 years. As we counted all cases on a weekly interval, the first week included 8 days since there was no incidence on January 1st.

The out-break data for each year was smoothed to eliminate weekend and holiday effects. We used Friedman's SuperSmoother ("supsmu" function in R program (R-CRAN, 2014)), a symmetric k-nearest neighbor linear least squares fitting procedure, with varying bandwidth selected by local cross validation (Friedman, 1984).

2.2. The probability density function of the incubation period

Different incubation periods by region have been reported (Contacos *et al.*, 1972; Garnham *et al.*, 1975; Adak *et al.*, 1998). P. vivax from temperate countries tend to have a longer incubation time than from tropical countries; however, some tropical malaria have long incubation period (Mangoni *et al.*,

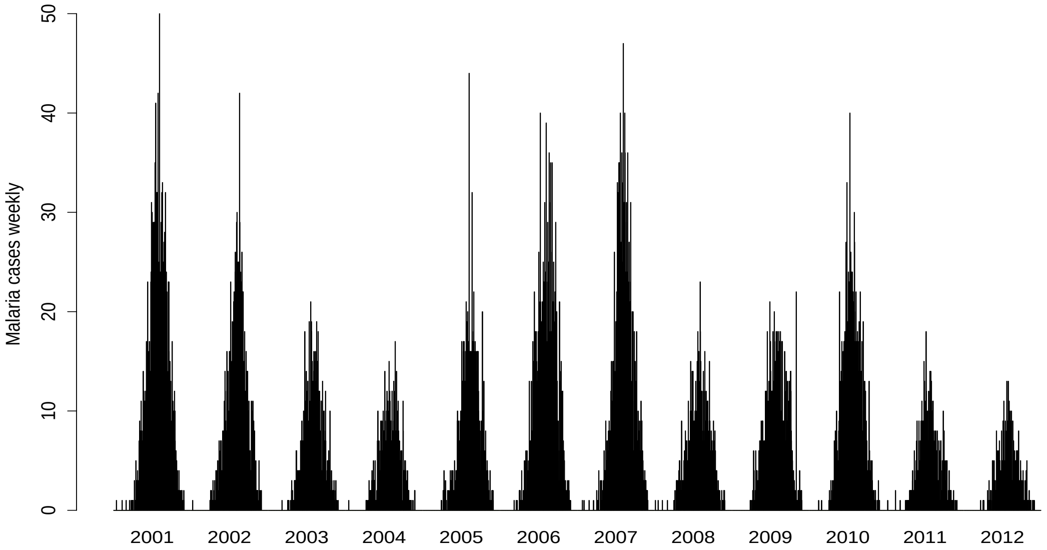

Figure 1: *Reported cases of malaria per week in Korea between 2001 and 2012. Note the characteristic cyclicity and slightly decreasing tendency.*

2003).

The incubation period of Plasmodium vivax in South Korea was investigated by Nishiura *et al.* (2007). They concluded that the incubation period of P. vivax in South Korea, consisted of short and long incubation periods. A total of 142 cases(63.1%) out of 225 with short incubation periods were fitted with a gamma distribution, $\Gamma(1.2, 22.2)$, and 83 cases(36.9%) with long incubation periods were fitted with a normal distribution, $N(337.4, 40.6^2)$.

Suppose a random variable T denotes the day which an infected man actually starts his clinical course (with the infection at day 0). So T is same as the incubation period. Then the probability density function(pdf) of T is represented by a mixture of two distributions;

$$f_T(d) \sim 0.63\Gamma(1.2, 22.2) + 0.37N\left(337.4, 40.6^2\right). \tag{2.1}$$

Figure 2 shows the probability density function of the incubation period estimated by Nishiura *et al.* (2007).

The long incubation period P. vivax infection is due to the hypnozoites stage. In the hypnozoites stage, the sporozoites is discharged from the salivary glands of the hibernating mosquito in the hepatic cell without multiplication (Cogswell, 1992).

2.3. Prevalence of malaria

To calculate the prevalence of malaria at week w of year y, we need to consider it through three years because the incubation period lasts up to 104 weeks(two years). We will derive prevalence for each years, and add up for three years. Denote P_w^y is the prevalence at week w of year y, for $w = 1, \ldots, 52$, G_t^y is the estimated numbers of infection at week t of year y, and S_t is the survival function of the incubation period T at week t.

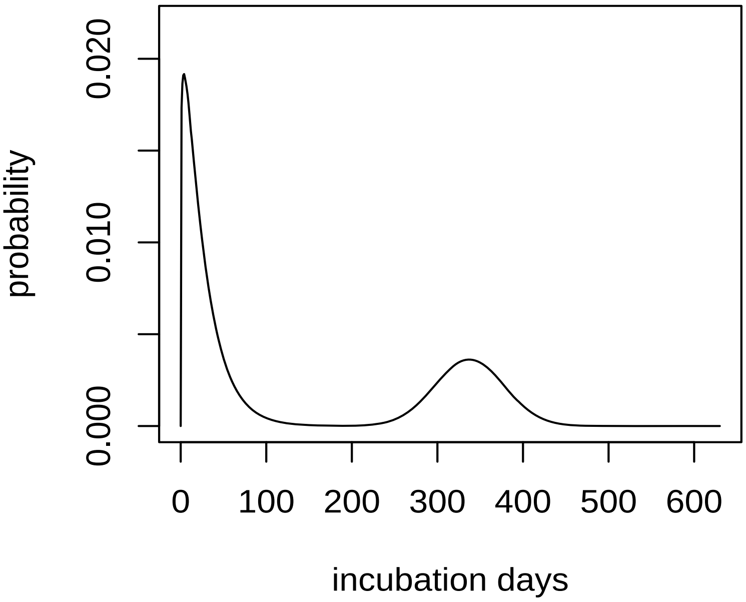

Figure 2: *The probability density function of incubation period of P. vivax in South Korea estimated by Nishiura et al. (2007).*

Note that the survival function is

$$S_t = \Pr(T > t) = 1 - F_T(t) = 1 - \sum_{w=1}^{t} f_w,$$

where $F_T(t)$ is the cumulative distribution function of the incubation period T, and f_w is the incubation probability for each week w. Here f_w is computed by adding the probabilities given in (2.1) for seven days in the week w. S_t means the probability that an infected man is still in the incubation period after the time t.

The prevalence P_w^y who have infected between the week w of year $y-2$ and the last (52rd) week of year $y-2$ is obtained by the equation, $\sum_{t=w}^{52} G_t^{y-2} S_{104-(t-w)}$. The prevalence P_w^y who have infected at year $y-1$ is obtained by the equation, $\sum_{t=1}^{52} G_t^{y-1} S_{52+w-t}$. The prevalence P_w^y who have infected between the first week of year y and the current week of year y is obtained by the equation, $\sum_{t=1}^{w-1} G_t^y S_{w-t}$.

Since we assume that $G_t^{y-2} = G_t^{y-1} = G_t^y = G_t$, we obtain the prevalence at week w of year y by adding the above three equations with G_t as:

$$P_w = \sum_{t=w}^{52} G_t S_{104-(t-w)} + \sum_{t=1}^{52} G_t S_{52+w-t} + \sum_{t=1}^{w-1} G_t S_{w-t}, \tag{2.2}$$

for $w = 1, \ldots, 52$.

The computed prevalence are drawn (solid line of the upper part) in Figure 4. Here G_t is obtained by using the back-calculation formula and maximum likelihood estimation method. The details of calculating G_t are described in the Appendix. Table 1 indicates the values of G_t estimated from data of South Korea. The similar approach is also provided in Lee *et al.* (2014).

Table 1: Results obtained by the proposed method utilizing back-calculations for 52 weeks: The estimated weekly number of infections and the prevalence with 95% confidence interval in parenthesis.

Week	Estimated infection numbers	Prevalence of malaria	Week	Estimated infection numbers	Prevalence of malaria
1	0	534 (503, 560)	27	108	516 (467, 542)
2	0	533 (503, 559)	28	113	553 (501, 582)
3	0	533 (502, 559)	29	115	590 (535, 621)
4	0	532 (502, 558)	30	111	625 (570, 657)
5	0	531 (501, 557)	31	104	655 (597, 686)
6	0	531 (500, 556)	32	96	677 (622, 710)
7	0	530 (499, 554)	33	85	693 (638, 726)
8	0	528 (498, 553)	34	71	701 (646, 733)
9	0	526 (496, 551)	35	56	701 (648, 732)
10	0	523 (493, 549)	36	45	692 (642, 726)
11	0	520 (490, 545)	37	35	680 (630, 712)
12	0	515 (486, 540)	38	27	665 (618, 697)
13	0	509 (481, 533)	39	22	649 (605, 679)
14	0	501 (472, 525)	40	18	634 (592, 664)
15	0	492 (463, 515)	41	12	622 (581, 651)
16	0	480 (450, 502)	42	0	609 (570, 637)
17	0	467 (436, 489)	43	0	589 (553, 616)
18	0	451 (419, 472)	44	0	574 (539, 601)
19	4	432 (400, 454)	45	0	563 (529, 590)
20	20	415 (382, 437)	46	0	555 (522, 581)
21	30	410 (376, 433)	47	0	549 (516, 575)
22	43	410 (373, 433)	48	0	544 (512, 570)
23	56	416 (378, 440)	49	0	541 (510, 567)
24	71	429 (388, 454)	50	0	538 (507, 564)
25	87	450 (405, 475)	51	0	537 (505, 563)
26	100	480 (434, 508)	52	0	535 (504, 561)

The 95% confidence intervals for the prevalence were calculated by using the parametric bootstrap method. By noting that there are two components(G_w and S_k) in (2.2), we considered the variations due to G_w and S_k. The reason why we deal with the variation due to S_k is that the pdf (2.1) of the incubation period is an estimated function and there is uncertainty. The algorithm to compute the confidence intervals using the bootstrap method is given as follows. Firstly, for G_w, we assumed that G_w follows a Poisson distribution with mean $\widehat{G_w}$, the estimated infection numbers for the week w as obtained in the Appendix. So a Poisson random variate with mean $\widehat{G_w}$ is generated for each week w, where $w = 1, 2, \ldots, 52$. Secondly, for S_k, 730 random numbers from the pdf (2.1) of the incubation period are generated. Then the empirical survival function $(\widehat{S_k})$ is constructed from 730 random numbers, by adding those of the corresponding seven days for the week k, where $k = 1, 2, \ldots, 114$. Using the generated G_w and the computed $\widehat{S_k}$, we can calculate the prevalence by (2.2). This consists one series of the prevalence for a year. We repeat this procedure B times to get the B series of the prevalence. Then, the $100 \times (1-\alpha)\%$ confidence interval of the prevalence at a week w is obtained as:

$$\left(P_{(B\frac{\alpha}{2})}, P_{B(1-\frac{\alpha}{2})}\right), \tag{2.3}$$

where $P_{(B\frac{\alpha}{2})}$ is the $B \times (\alpha/2)$-th ascending order statistic among the B bootstrap samples of the prevalence at a fixed week w. This confidence interval construction for a fixed week w is gone through for every week w, for $w = 1, 2, \ldots, 52$. Here, we used $B = 1000$ and $\alpha = 0.05$ for actual computation. Figure 4 shows the estimated confidence intervals (dotted lines of the upper part).

3. Probabilities of Reactivation and Parasitemia

In the hypnozoites state, malaria hibernates in hepatic cells and is not present in the bloodstream before reactivation. Thus, we calculate the transmittable prevalence that is the number of transmittable persons with parasitemia in the incubation period.

3.1. Reactivation distribution

To compute the transmittable prevalence(TP), we need to know the probability density of parasitemia. Infections with long incubation periods actually start their clinical course after reactivation. It is almost impossible for every infection with a long incubation period to reactivate on the same day. Therefore, it is reasonable to think that the normal distribution of a long incubation period is contributed to by the distribution of the reactivation time. We assume that the reactivation time distribution follows a normal distribution and the probability of the incubation time after reactivation is the same as the probability of a short-term incubation. The latter assumption means that the incubation time for each reactivation at the long-incubation period follows the Gamma (1.2, 22.2) distribution, which is the probability of a short-term incubation. So we can infer that the probability of a long-term incubation is actually the convolution of a reactivation and a short-term incubation. This convolution is represented by the following Equation (3.1).

Since transmission is possible after reactivation, we need to know a distribution of reactivation time first and we assume that reactivation time follows a normal $N(\mu_0,\ \sigma_0^2)$ distribution. This normality assumption is reasonable because the long-term incubation period follows a normal distribution. Here, the parameters $\mu_0,\ \sigma_0$ are unknown and to be estimated. It is estimated by using the probability density function of the incubation period, the back-calculation formula and the least squares method. Instead of assuming that N_d follows a normal distribution, one can take a distribution-free way. It is discussed in the Section 5.

Denote N_d as the probability of reactivation at day d in the incubation. Let L_d be the probability at day d which follows the normal $N(337.4,\ 40.6^2)$ distribution. Let p_d be the probability at day d which follows the Gamma (1.2, 22.2) distribution. Here L_d and p_d are the probabilities of day d long-term and short-term incubation, respectively, which distributions are came from Nishiura *et al.* (2007) as given in (2.1). Then one estimation of L_d, denoting by $\widehat{L_d}$, is calculated using the following back-calculation formula:

$$\widehat{L_d} = \sum_{k=1}^{320} N_{d-k} p_k, \tag{3.1}$$

for $d = 337 - 4 \times 40, \ldots, 337 + 4 \times 40$, and $N_{d-k} = 0$ for $d \leq k$. Here, the range of d is chosen as $[\mu - 4\sigma,\ \mu + 4\sigma]$, where this interval covers 99.9% of non-zero probabilities of $N(\mu,\ \sigma^2)$ distribution.

Now, parameters μ_0 and σ_0 are estimated by minimizing the following sum of squared differences:

$$Q\left(\mu_0, \sigma_0^2\right) = \sum_{d=177}^{497} \left(L_d - \widehat{L_d}\right)^2. \tag{3.2}$$

This optimization was performed numerically by using a routine (“optim”) in R program (R-CRAN, 2014). This routine utilizes the quasi-Newton algorithm. A finite-difference approximation was used for the gradient of the objective function. We set several initial values for $(\mu_0,\ \sigma_0)$, and treated the best local minimizer as the global minimizer.

Table 2: Calculated values of the probability of transmission at day d, obtained by the Equation (3.4). The notes in parenthesis, that is, ‘start’ and ‘peak’ denotes the starting time and the peak time of the probability of transmission which is obtained based on only the long-term incubation period.

Day	1	5	10	15	20	25	30	35
Probability	0.6175	0.5462	0.4576	0.3791	0.3119	0.2556	0.2087	0.1700
Day	40	45	50	55	60	80	100	120
Probability	0.1382	0.1122	0.0909	0.0736	0.0596	0.0253	0.0107	0.0045
Day	150	180	210	240(start)	250	260	270	280
Probability	0.0012	0.0003	0.0004	0.0041	0.0079	0.0140	0.0231	0.0352
Day	290	300	310	320	333(peak)	350	360	370
Probability	0.0498	0.0653	0.0796	0.0903	0.0960	0.0879	0.0769	0.0637
Day	380	390	400	410	420	430	440	450
Probability	0.0501	0.0377	0.0273	0.0192	0.0131	0.0088	0.0058	0.0038
Day	460	500						
Probability	0.0025	0.0000						

The actual estimated values of parameters in a normally distributed reactivation time are $\hat{\mu}_0 = 313.465$, and $\hat{\sigma}_0 = 34.503$. From these values, we know that the reactivation starts 24 days earlier, on average, than the mean of the long-term incubation period. The standard deviation of the reactivation distribution is about 85% smaller than the long-term incubation period.

3.2. The probability of transmission

As the survival function was used importantly in calculating the prevalence of malaria in subsection 2.3, we would like to derive the survival function of the incubation period under the transmittable condition (with parasitemia) to compute the TP. Let t_0 be the possible start day of the reactivation for a patient with a long term incubation period. Actually we set $t_0 = 337 - 5 \times 40 = 137$. Let T^l denotes the long-term incubation period under transmittable condition. Then the probability of transmission at day d for the long-term incubation period ($d \geq t_0$) is

$$\Pr\left(T^l > d\right) = \sum_{k=0}^{d-t_0} N_{d-k} S_k^0, \quad \text{for } d \geq t_0, \tag{3.3}$$

where S_k^0 is the survival function of the day k with the short term incubation period, and N_d is the probability of the reactivation at day d which follows a normal distribution with mean 314.465 and standard deviation 34.503 as obtained in the above subsection.

Now the probability of transmission at the day d is defined as a mixture of two probability functions of the short-term and of long-term incubation periods;

$$S_d^p = 0.631 \times S_d^0 + 0.369 \times \Pr\left(T^l > d\right). \tag{3.4}$$

This is used to compute the TP. Note that malaria is always transmittable during the short-term incubation period. Thus the short-term survival function (S_d^0) is used without modification.

Figure 3 shows this function. The maximum value 0.096 for the long-term incubation period occurs at the 333rd day. Table 2 provides the selected numerical values of this function.

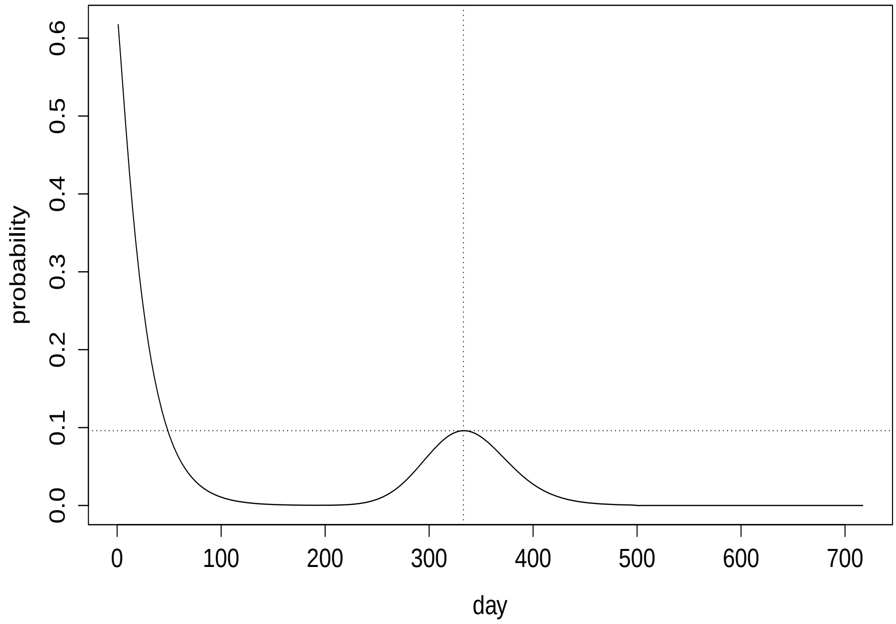

Figure 3: *A mixture of the probability functions of short-term and long-term incubation periods with parasitemia, obtained by the Equation (3.4). The maximum probability 0.096 for the long-term incubation period occurs at the 333rd day.*

4. Transmittable Prevalence

As an analogy from the Equation (2.2), we compute the transmittable prevalence(TP) using the following convolution;

$$P_w^t = \sum_{k=w}^{52} G_k S^p_{104-(k-w)} + \sum_{k=1}^{52} G_k S^p_{52+w-k} + \sum_{k=1}^{w-1} G_k S^p_{w-k}, \tag{4.1}$$

for $w = 1, \ldots, 52$, where P_w^t is the TP at week w, G_k is the infected numbers at week k, and S_k^p is the probability of transmission at week k computed from Equation (3.4). Here, S_k^p for week k is calculated by adding the corresponding daily values for seven days. This calculation is done over two years because some S_k^p are non-zero over 104 weeks. Figure 4 shows the result of the TP (solid line of the lower part) and its 95% confidence intervals (dotted lines of the lower part) obtained from data of South Korea. Table 3 provides the numbers corresponding to this figure.

The 95% confidence intervals for the TP were calculated using the parametric bootstrap method. By noting that there are two components (G_k and S_w^p) in (4.1), we considered the variations due to G_k and S_w^p. The below is the algorithm to compute the confidence intervals using bootstrap method. Firstly, for G_k, a Poisson random variate with mean $\widehat{G_k}$ is generated for each week k, where $k = 1, 2, \ldots, 52$, as we did in the Subsection 2.3. Secondly, for S_w^p, a normal variate with mean p_w and variance $p_w(1 - p_w)$ is generated for each week w, where $w = 1, 2, \ldots, 104$. This computation is repeated for every k and w. Then, using generated G_k and S_w^p's for every week k and w, we can calculate the TP by (4.1). This constructs one series of the TP for a year. We repeat this procedure B times to get the B series of the TP. Then, the $100 \times (1 - \alpha)\%$ confidence interval of the TP at a week

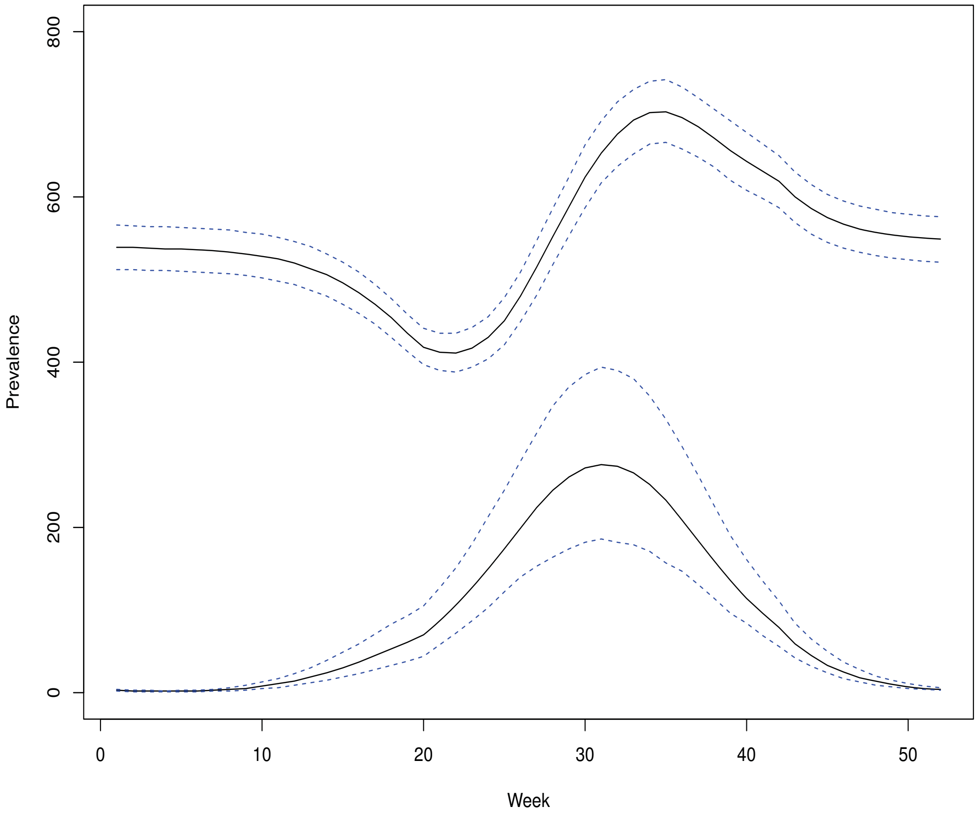

Figure 4: *Distributions of prevalence (solid line of the upper part) and the transmittable prevalence (solid line of the lower part) of malaria in Korea. 95% confidence intervals are drawn by dotted lines for each prevalence. The maximum (276 cases) of the transmittable prevalence occurs at the 31st week.*

k is obtained as;

$$\left(\mathrm{TP}_{(B\frac{\alpha}{2})}, \mathrm{TP}_{B(1-\frac{\alpha}{2})}\right), \tag{4.2}$$

where $\mathrm{TP}_{(B\frac{\alpha}{2})}$ is the $B \times (\alpha/2)$-th ascending order statistic among the B bootstrap samples of the TP at a fixed week k. This confidence interval construction for a fixed week k is gone through for every week k, for $k = 1, 2, \ldots, 52$. Here, we used $B = 1000$ and $\alpha = 0.05$ for actual computation.

From Figure 4, we can see the TP is distributed similarly as a normal distribution, with a shape that looks like the smoothed incidence distribution. The maximum (276 cases) occurs at the 31$^{\text{st}}$ week. The cases in incubation with parasitemia are less than the naive prevalence (solid line of the upper part). The TP is about 60% reduction at the peak from the naive prevalence. The width of confidence intervals of the TP is proportional to the height of the TP. The largest width of confidence interval at the peak of the TP is 206. It is possible to induce the maximum standard error of the TP be about 52, assuming the normal distribution.

It is notable that the TP between the 50$^{\text{th}}$ week and the 14$^{\text{th}}$ week is very small. Thus, we may consider easing the restriction on malaria for blood donations during this period.

Table 3: Results obtained by the proposed method utilizing back-calculations for 52 weeks: The estimated weekly number of the transmittable prevalence and 95% confidence interval in parenthesis.

Week	Transmittable prevalence	Week	Transmittable prevalence
1	3 (2,4)	27	224 (153,313)
2	2 (1,3)	28	245 (164,347)
3	2 (1,3)	29	261 (174,369)
4	2 (1,2)	30	272 (182,384)
5	2 (1,3)	31	276 (186,392)
6	2 (1,3)	32	274 (182,388)
7	3 (2,4)	33	266 (179,380)
8	4 (2,6)	34	252 (171,358)
9	5 (3,9)	35	233 (157,329)
10	8 (5,13)	36	209 (147,297)
11	11 (6,17)	37	184 (131,262)
12	14 (9,23)	38	159 (114,226)
13	19 (12,30)	39	136 (96,189)
14	24 (15,39)	40	114 (84,160)
15	30 (19,48)	41	96 (69,134)
16	37 (23,58)	42	79 (56,111)
17	45 (28,71)	43	59 (42,84)
18	53 (33,82)	44	45 (32,65)
19	61 (38,93)	45	33 (24,49)
20	70 (44,104)	46	25 (17,37)
21	87 (58,127)	47	18 (13,27)
22	106 (72,151)	48	14 (9,20)
23	127 (87,180)	49	10 (7,15)
24	150 (103,213)	50	7 (5,11)
25	174 (122,244)	51	5 (4,8)
26	199 (140,280)	52	4 (3,6)

5. Discussion

We analyzed the incidence data on a weekly basis even though the original source from the KCDC was on a daily basis. The weekly data was then smoothed to eliminate weekend and holiday effects. The first time in our study, we tried to calculate the daily infection rate using the daily incidence data, but it was very difficult because there were too many regression coefficients ($n = 365$). The variation of the daily infection rate was too big to accept when we calculated the rate using the matrix inversion method. We also did not use 2 weeks of interval data because the loss of information was considerable.

Transmittable prevalence, the number of transmittable people with parasitemia if the disease can be transmitted directly, is concentrated in the summer with a peak at the 31st week. In malaria situations, it is possible for people with malaria to donate blood to people needing blood transfusions (Kitchen and Chiodini, 2006; Mary *et al.*, 2001). In South Korea, any visitor to a malarial endemic area is not permitted to donate blood for 1 year. We may consider easing the restriction based on estimated infection rates and transmittable prevalence.

We developed a new statistical methodology tp estimate the transmittable prevalence associated with short- and long-term incubation periods. We defined a new probability function of incubation period with parasitemia. We obtained probabilities of reactivation and of parasitemia by repeatedly using the back-calculation formula. Reviewers recommended a performance evaluation of our estimation method for the transmittable prevalence and of bootstrap approach for constructing confidence intervals. They also suggested a simulation study; however, it is hard and left as a topic for future study.

A normal distribution is assumed for the reactivation time distribution when estimating transmittable prevalence. This parametric assumption can be removed. We may estimate the distribution directly by the least squares method without assuming a parametric distribution, where the reactivation probabilities are treated as regression coefficients; consequently, some statistical considerations and numerical techniques may be required.

We calculated the confidence intervals for the prevalence by using a parametric bootstrap in Subsection 2.3. One can try a nonparametric bootstrap approach as the below. For this purpose, we treat a time series of malaria incidences of each year as an observation, so that consisted of 13 observations. We construct a bootstrap sample from these 13 time series by sampling with replacement. From this bootstrap sample, we estimate G_w by minimizing the Equation (A.3), and denote it $\widehat{G}_w^{(1)}$. This procedure is repeated B times to construct $\widehat{G}_w^{(1)}, \ldots, \widehat{G}_w^{(B)}$ for every w. Using $\widehat{G}_w^{(i)}$, we can calculate the B series of prevalence by using (2.2). Then, the $100 \times (1-\alpha)\%$ confidence interval of the prevalence at a week w is obtained by (4.2). This confidence interval construction for a fixed week w is gone through for every week w, for $w = 1, 2, \ldots, 52$. This nonparametric bootstrap approach might be more honest to data than the parametric bootstrap (presented in the Subsection 2.3) that requires a numerical optimization routine to estimate G_w by minimizing the Equation (A.3) for each bootstrap sample, which is sometimes unsuccessful in reaching to the global minimizer. In addition, the computing time of this approach is more than that of a parametric bootstrap. This discussion on the confidence intervals for prevalence can be similarly applied to the transmittable prevalence after modification; in addition, one can try a semi-parametric bootstrap method (Kim, 2011) which combines one part of parametric bootstrap and the other part of nonparametric bootstrap.

We used the result of Nishiura *et al.* (2007) for the incubation period of P. vivax for what is essential for the back-calculation of infection rates. Kim *et al.* (2013) obtained the similar results with a shorter mean incubation period than Nishiura *et al.* (2007). An estimate of transmittable prevalence with the incubation periods of Kim *et al.* (2013) will change the results. This possibility is partly covered in the confidence interval computations in Sections 2.3 and 4 by considering the uncertainty of the survival functions that originated from the estimated pdf of Nishiura *et al.* (2007).

We think the malaria data of other countries can be analyzed in the same way as presented here if they have information about incubation periods for own malaria and incidence surveillance data. For example, it can be applicable to some tropical malaria with long incubation periods (Mangoni *et al.*, 2003). This method can also be used for other infectious diseases.

6. Summary

We developed a new statistical methodology for estimating the transmittable prevalence associated with short-term and long-term incubation periods. The method was applied to data in South Korea. We computed the probabilities of reactivation and of parasitemia. Transmittable prevalence, the number of transmittable persons with parasitemia if the disease can be transmitted directly, is concentrated in summer with 276 a peak at the 31st week in South Korea. The transmittable prevalence is about 60% reduction at the peak from the naive prevalence. Estimated transmittable prevalence may be useful in modifying blood donation regulations related to malaria.

Appendix: Estimating infection distribution by the back-calculation

Assuming that the out-break observations follow a Poisson distribution, we can estimate the infection distribution using the back-calculation formula and the maximum likelihood method.

From the back-calculation method (Brookmeyer and Gail, 1988; Bacchetti *et al.*, 1993; De Angelis *et al.*, 2004), we have for $y = 2001, \ldots, 2012$ and for $w = 1, \ldots, 52$,

$$A_w^y = \sum_{k=0}^{103} f_{104-k} G_{w+k} + \varepsilon, \tag{A.1}$$

where f_w is the incubation probability computed for each week w, G_w is the infection numbers for the week w, A_w^y is a random variable representing the malaria cases at y year and w week, and the random variable ε is the error term. The range of k (from 0 to 103) is set to cover two years. Thus, we actually assume that A_w^y follows a Poisson distribution with a mean function

$$\lambda_w = \sum_{k=0}^{103} f_{104-k} G_{w+k}. \tag{A.2}$$

Since we already know A_w^y and f_w, the unknown quantities G_{w+k} are treated as regression coefficients and are subject to being estimated. Here, we assume that $G_w = G_{w+52} = G_{w+104}$ for $w = 1, \ldots, 52$. f_w is computed by adding the corresponding daily incubation probabilities for seven days.

The log-likelihood function of G_w for given data $\hat{A}_w^y$ is proportional to

$$\sum_{y=2001}^{2012} \sum_{w=1}^{52} \left[\hat{A}_w^y \log \lambda_w - \lambda_w \right], \tag{A.3}$$

where $\hat{A}_w^y$ is the supersmoothed value from the observed malaria cases at y year and w week, and λ_w is the mean function of Equation (A.2). Since no explicit maximizers of Equation (A.3) exist, a numerical optimization routine is needed to estimate G_w for $w = 1, \ldots, 52$. The estimator is denoted as $\widehat{G_w}$. We use quasi-Newton algorithm ("optim" function) in R program (R-CRAN, 2014) to minimize the negative value of the Equation (A.3). The estimated infection numbers are provided in Table 1.

Acknowledgments

The authors thank Seok Ju Park who helped with computations.

References

Adak, T., Sharma, V. P. and Orlov, V. S. (1998). Studies on the Plasmodium vivax relapse pattern in India, *American Journal of Tropical Medicine and Hygiene*, **59**, 175–179.

Bacchetti, P., Segal, M. and Jewell, N. P. (1993). Back-calculation of HIV infection rate (with discussion), *Statistical Science*, **8**, 82–119.

Brookmeyer, R. and Gail, M. H. (1988). A method of obtaining short-term projections and lower bounds on the size of the AIDS epidemic, *Journal of the American Statistical Association*, **83**, 301–308.

Burket, D. A., Lee, W. J., Lee, K. W., Kim, H. C., Lee, H. I., Lee, J. S., Shin, E. H., Wirtz, R. A., Cho, H. W., Claborn, D. M., Coleman, R. E., Kim, W. Y. and Klein, T. A. (2002). Late season commercial mosquito trap and host seeking activity evaluation against mosquitoes in a malarious area of the Republic of Korea, *Korean Journal of Parasitology*, **40**, 45–54.

Centers for Disease Control and Prevention; CDC (2006). Available from: http://www.cdc.gov/malaria/

Cogswell, F. B. (1992). The hypnozoite and relapse in primate malaria, *Clinical Microbiology Reviews*, **5**, 26–35.

Contacos, P. G., Collins, W. E., Jeffery, G. M., Krotoski, W. A. and Howard, W. A. (1972). Studies on the characterization of Plasmodium vivax strains from Central America, *American Journal of Tropical Medicine and Hygiene*, **21**, 707–712.

De Angelis, D., Hickman, M. and Yang, S. (2004). Estimating long-term trends in the incidence and prevalence of opiate use/injecting drug use and the number of former users: Back-calculation methods and opiate overdose deaths, *American Journal of Epidemiology*, **160**, 994–1004.

Friedman, J. H. (1984). *A Variable Span Scatterplot Smoother*, Technical Report No.5, Laboratory for computational statistics, Stanford University.

Garnham, P. C. C., Bray, R. S., Bruce-Chwatt, L. J., Draper, C. C., Killick-Kendrick, R., Sergiev, P. G., Tiburskaja, N. A., Shute, P. G. and Maryon, M. (1975). A strain of Plasmodium vivax characterized by prolonged incubation: Morphological and biological characteristics, *Bulletin of the World Health Organization*, **52**, 21–32.

Goubar, A., Bitar, D., Cao, W. C., Feng, D., Fang, L. Q. and Desenclos, J. C. (2009). An approach to estimate the number of SARS cases imported by international air travel, *Epidemiology & Infection*, **137**, 1019–1031.

Greenwood, B. and Mutabingwa, T. (2002). Malaria in 2002, *Nature*, **415**, 670–672.

Hall, H. I., Song, R., Rhodes, P., Prejean, J., An, Q., Lee, L. M., Karon, J., Brookmeyer, B., Kaplan, E. H., McKenna, M. T., Janssen, R. S. and for the HIV Incidence Surveillance Group (2008). Estimation of HIV incidence in the United States, *Journal of the American Medical Association*, **300**, 520–529.

Kim, J. H. (2011). Semi-parametric bootstrap confidence intervals for high-quantiles of heavy-tailed distributions, *Communications for Statistical Applications and Methods*, **18**, 717–732.

Kim, S. J., Kim, S. H., Jo, S. N., Gwack, J., Youn, S. K. and Jang, J. Y. (2013). The long and short incubation periods of plasmodium *vivax* malaria in Korea: The characteristics and relating factors, *Infection & Chemotherapy*, **45**, 184–193.

Kitchen, A. and Chiodini, P. (2006). Malaria and blood transfusion, *Vox Sanguinis*, **90**, 77–84.

Korean Center for Disease Control (2007). *Malaria Infection Control and Management Policy*, KCDC, Seoul.

Korean Center for Disease Control (2008). *Statistics of Communicable Diseases*, Seoul: KCDC, Available from: http://stat.cdc.go.kr.

Law, M., Lynskey, M., Ross, J. and Hall, W. (2001). Back-projection estimates of the number of dependent heroin users in Australia, *Addiction*, **96**, 433–443.

Lee, J. S., Lee, W. J., Cho, S. H. and Ree, H. I. (2002). Outbreak of vivax malaria in areas adjacent to the demilitarized zone, South Korea, *American Journal of Tropical Medicine and Hygiene*, **66**, 13–17.

Lee, Y., Jang, H., Rhee, J. A. and Park, J. S. (2014). Statistical estimations for Plasmodium vivax malaria in South Korea, *Asia Pacific Journal of Tropical Medicine*, in press.

Mangoni, E. D., Severini, C., Menegon, M., Romi, R., Ruggiero, G. and Majori, G. (2003). Case report: An unusual late relapse of Plasmodium vivax malaria, *American Journal of Tropical Medicine and Hygiene*, **68**, 159–160.

Mary, M., Gary, T., Mary, C. and Monica, P. (2001). Transfusion-transmitted malaria in the United States from 1963 through 1999, *New England Journal of Medicine*, **344**, 1973–1978.

Mezzetti, M. and Robertson, C. (1999). A hierarchical Bayesian approach to age-specific backcalculation of cancer incidence rates, *Statistics in Medicine*, **18**, 919–933.

Nishiura, H., Lee, H. W., Cho, S. H., Lee, W. G., In, T. S., Moon, S. U., Chung, G. T. and Kim, T. S. (2007). Estimates of short- and long-term incubation periods of Plasmodium vivax malaria in the Republic of Korea, *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **101**, 338–343.

Park, J. W., Klein, T. A., Lee, H. C., Pacha, L. A., Ryu, S. H., Yeom, J. S., Moon, S. H., Kim, T. S., Chai, J. Y., Oh, M. D. and Choe, K. W. (2005). Vivax malaria: A continuing health threat to the Republic of Korea, *American Journal of Tropical Medicine and Hygiene*, **69**, 159–167.

Punyacharoesin, N. and Viwatwongkasem, C. (2009). Trends in three decades of HIV/AIDS epidemic in Thailand by nonparametric backcalculation method, *AIDS*, **23**, 1143–1152.

R-CRAN (2014). R program, Available from: http://www.r-project.org

Ree, H. I., Hwang, U. W., Lee, I. Y. and Kim, T. E. (2001). Daily survival and human blood index of Anopheles sinensis, the vector species of malaria in Korea, *American Mosquito Control Association*, **17**, 67–72.

Rogers, D. J. and Randolph, S. E. (2000). The global spread of malaria in a future, warmer world, *Science*, **289**, 1763–1766.

WHO (2005). *World Malaria Report*.

Received May 20, 2014; Revised July 20, 2014; Accepted August 24, 2014