Mathematical and Statistical Approaches in Epidemiological Investigation of Hospital Infection: A Case Study of the 2015 Middle East Respiratory Syndrome Outbreak in Korea

Youngsuk Ko¹ and Eunok Jung²*

1) Institute of Mathematical Sciences, Konkuk University, Korea

2) Department of Mathematics, Konkuk University, Korea

*junge@konkuk.ac.kr
Summary

Introduction

Epidemiological investigations are used to identify outbreaks, collect data, and implement control measures, playing a crucial role in disease control and prevention. Mathematical and statistical approaches enhance these investigations by incorporating data analysis to understand disease characteristics and provide insights.

Methods

To estimate the transmissibility of a disease, we introduce a process to derive the likelihood function using individual patient information from epidemiological investigation. This method was applied to the case of Pyeongtaek St. Mary's Hospital during the 2015 Middle East Respiratory Syndrome outbreak in Korea. A stochastic model was developed, and scenario analysis reflecting actual outbreak progress, risk factors, and mask mandates was conducted.

Results

We applied transmission rate estimation to the Pyeongtaek St. Mary's Hospital case, showing a high patient-to-patient transmission rate. The superspreader was observed to have approximately 25 times higher transmissibility than other patients. Given these conditions, if hospital transmission period is prolonged, number of cases could be three times higher than the actual incidence. The effect of mask-wearing in hospital was investigated based on the type of mask and the intensity of the intervention. It was found that the scale of epidemic could be reduced by a maximum of 77% and a minimum of 17%.

Conclusions

Through the application of mathematical and statistical methodologies in epidemiological investigations, this study identified and quantified risk factors. Methodology of this study can be easily adapted and applied to other diseases and is expected to help in establishing effective strategies to fight against emerging infectious diseases.
Introduction

Epidemiological investigations play a pivotal role in public health and serve as the backbone of disease control and prevention (1–4). They are used to uncover the patterns, causes, and effects of health and disease conditions in defined populations. The primary purpose of these investigations is to identify the occurrence of an outbreak or unusual increase in number of cases. This is followed by the collection of data pertaining to cases, their characteristics, and potential risk factors. These data are then used to implement control measures to prevent further spread of the disease and to forestall future outbreaks.

Mathematical and statistical approaches enhance the significance of epidemiological investigations. Information aggregated through these investigations can be processed to understand the unique characteristics of specific infectious diseases, such as their incubation or latent periods, and presented as statistical distributions to aid future research (5–7). For example, during the COVID-19 pandemic, epidemiological information from individual patients was processed on a large scale using statistical approaches (8–11). These analyses revealed changes in contact patterns between age groups during COVID-19, which differed from those observed before the pandemic. This information was then used for short- and long-term predictions, considering factors such as vaccine prioritisation and policy decisions. By integrating epidemiological investigations with mathematical and statistical analyses, this approach provides evidence that enhances the reliability and accuracy of model results.

In this study, we introduce a process for interpreting epidemiological investigations using mathematical and statistical methods to estimate key parameters and transmission rates. As a case study, we apply this method to an outbreak at Pyeongtaek St. Mary's Hospital (PMH), the initial site of hospital transmission during the 2015 Middle East Respiratory Syndrome (MERS) outbreak in Korea, where 36 out of 186 total cases were identified (12–14). The spread in Korea was largely due to superspreaders, patients who caused secondary infections in more than six people (15). Of the 186 cases, it was suspected that 15 transmitted the disease to others, with five identified as superspreaders. To observe transmission rates in detail, we categorised individuals into superspreaders, healthcare workers (HCWs), patients, and visitors based on actual epidemiological investigations. We then developed a model that reflects these investigations and simulates a realistic outbreak. A scenario-based analysis was introduced to quantify risk factors such as the infectious
period of infected individuals, including superspreaders, and the effectiveness of mask mandates within the hospital. The methodology used in this study, which relatively easily transforms the situation of hospital infections into a likelihood function, can be applied to other hospitals or infectious disease outbreaks. We expect that these findings can significantly enhance both reactive and preventive measures.
Materials and Methods

Probabilities and likelihoods during the outbreak

Two necessary pieces of information expected from the epidemiological investigation are: (1) when were the infectious individuals suspected of transmitting the disease and (2) when were the infected individuals infected with the disease (or the time range in which they were exposed to the disease). These factors influence the probability of individuals becoming infected at certain time points. Let us consider a classic SIR model formulated using ordinary differential equations for a susceptible population (16). A susceptible host can be infected by an infectious host at a rate $\beta$. Let $N$ indicate the total number of hosts. When assuming frequency-dependent transmission, ignoring natural birth and death (17), and fixing the number of infectious hosts as $N_i$, the following equation is obtained:

$$\frac{dS}{dt} = -\beta \frac{S}{N} N_i$$

This equation transitions to a linear form, facilitating a straightforward resolution. Given the assumption that the initial $S$ is one to consider the status of a single host and $N$ is a constant, the equation is solved as follows:

$$S(t) = \exp\left(-\beta \frac{N_i}{N} t\right)$$

This solution represents the probability that the host remains in the susceptible state for up to time $t$. When considering a unit of time, defined as $t = 1$, the subsequent probability is conceptualised as follows:

$$q_S = \exp\left(-\beta \frac{N_i}{N}\right).$$

In this framework, $q_S$ is interpreted as the probability of a host maintaining an uninfected status over a unit of time. Conversely, the probability of a host becoming infected is represented as $q_I$, and expressed as follows:

$$q_I = 1 - q_S = 1 - \exp\left(-\beta \frac{N_i}{N}\right).$$

Let us consider a situation in which multiple hosts coexist, as shown in Figure 1. At time $t$, there are $N$ hosts in total, with $N_S$ susceptible and $N_I$ infectious hosts. After a unit of time ($t + 1$), if $X$ hosts are infected and $Y$ hosts are not infected ($X + Y = N$), using previously calculated probabilities, we can calculate the following likelihood function:
\[ L = \prod_{i=1}^{X} \exp \left( -\beta \frac{N_i}{N} \right) \times \prod_{i=1}^{Y} \left( 1 - \exp \left( -\beta \frac{N_i}{N} \right) \right). \]

Since the epidemiological investigation is based on actual incidences, the value of \( \beta \) that maximises this likelihood function can represent the reality. Therefore, our goal is to determine the value of \( \beta \) that maximises the likelihood \( L \). In this study, the Metropolis-Hastings algorithm was used to sample the parameters to find the value of \( \beta \) (18). This explanation describes progress in one unit of time, whereas the actual case application considers multiple unit time steps and is expressed as the product of all probabilities.

Figure 1. Derivation of the likelihood function for a unit timestep considering both unsuccessful and successful disease transmissions.
Application of maximum likelihood estimation (MLE) to MERS nosocomial spread in Korea and model formulation

Kim provided insights into individual hosts considering their types (HCWs, patients, and visitors), encompassing their anticipated exposure times, expected transmission times, isolation times, and the number of individuals present in PMH during the 2015 MERS outbreak in Korea (14). Let $\beta_{AB}$ denote the transmission rate, where the subscripts $A$ and $B$ indicate the infector and infectee types, respectively. Symbols $\Omega$, $H$, $P$, and $V$ as subscripts indicate the types of hosts as superspreaders, HCWs, patients, and hospital visitors, respectively. We classified individuals as either infected or uninfected. The group of infected and uninfected individuals are represented as $\Lambda_I$ and $\Lambda_S$, respectively. Let $D_i$ and $\tilde{D}_i$ be the identifiers for the types of hosts in $\Lambda_S$ and $\Lambda_I$, respectively, and $T$ be the discrete time points. Considering the number of infectors at a specific time $k$ is $I_j(k)$, where the subscript $j$ indicates the host type, the likelihood of hosts in $\Lambda_S$ and $\Lambda_I$ is derived as follows:

$$L_S = \prod_{i \in \Lambda_S} \left\{ \prod_{k \in T} \exp \left( -\sum_{j \in \{\Omega, H, P, V\}} \beta_{jD_i} \frac{I_j(k)}{N} \right) \right\}$$

$$L_I = \prod_{i \in \Lambda_I} \left\{ \prod_{k \in T} \left( 1 - \exp \left( -\sum_{j \in \{\Omega, H, P, V\}} \beta_{j\tilde{D}_i} \frac{I_j(k)}{N} \right) \right) \right\}$$

$$L(B) = L_S \times L_I$$

where $B = \{\beta_{HH}, \beta_{HP}, \beta_{HV}, \beta_{PH}, \beta_{PP}, \beta_{PV}, \beta_{VH}, \beta_{VP}, \beta_{VV}, \beta_{\Omega H}, \beta_{\Omega P}, \beta_{\Omega V}\}$.

Our goal is to find $B$ that maximises $L(B)$. Note that $\beta_{HH}$, $\beta_{HP}$, and $\beta_{HV}$ are set to zero because the epidemiological investigation showed that there was no contagious period involving HCWs in PMH. In addition, symmetric transmission between patients and visitors was assumed, i.e., $\beta_{PV} = \beta_{VP}$.

In this study, we developed a susceptible-exposed-infectious-recovered (SEIR)-type model to investigate nosocomial spread, and Figure 2 provides a visual representation of our model. There were five stages of disease progression: susceptible ($S$), exposed ($E$), infectious ($I$), isolated ($Q$), and recovered ($R$). The subscripts indicate the types of hosts introduced. Figure 2 illustrates the different types of reactions: the non-Markovian process (transition from disease exposure, dashed line with an arrow) and the Markovian process (solid line with an arrow). These processes are listed in Table 1. To incorporate the infectious
period of the hosts into the model simulation, we fitted the data assuming a gamma distribution for the samples using the built-in MATLAB function (19).

**Figure 2. Flow diagram of the MERS intra-hospital transmission model.**

\[
\begin{align*}
\Lambda_H &= \frac{\beta_{HH}I_H + \beta_{PH}I_P + \beta_{VH}I_V + \beta_{OH}I_{OH}}{N} \\
\Lambda_P &= \frac{\beta_{HP}I_H + \beta_{PP}I_P + \beta_{VP}I_V + \beta_{OP}I_O}{N} \\
\Lambda_V &= \frac{\beta_{HV}I_H + \beta_{PV}I_P + \beta_{VV}I_V + \beta_{OV}I_O}{N}
\end{align*}
\]
Table 1: Description of reactions in the MERS intra-hospital transmission model.

<table>
<thead>
<tr>
<th>Event</th>
<th>Reaction</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infection of HCW</td>
<td>Markovian</td>
<td>Propensity: $S_H \beta_{HH} I_H + \beta_{PH} I_P + \beta_{VH} I_V + \frac{\beta_{HH} I_H}{N}$</td>
<td>Fitted</td>
</tr>
<tr>
<td>Infection of patient</td>
<td>Markovian</td>
<td>Propensity: $S_P \beta_{HP} I_H + \beta_{PP} I_P + \beta_{VP} I_V + \frac{\beta_{AP} I_P}{N}$</td>
<td>Fitted</td>
</tr>
<tr>
<td>Infection of visitor</td>
<td>Markovian</td>
<td>Propensity: $S_V \beta_{HV} I_H + \beta_{PV} I_P + \beta_{VH} I_V + \frac{\beta_{AV} I_V}{N}$</td>
<td>Fitted</td>
</tr>
<tr>
<td>Symptom onset</td>
<td>Non-Markovian</td>
<td>Delay: gamma distribution, Mean: 6.99, SD: 3.31</td>
<td>Fitted, (20)</td>
</tr>
<tr>
<td>Isolation</td>
<td>Non-Markovian</td>
<td>Delay: gamma distribution, Mean: 5.53, SD: 3.70</td>
<td>Fitted, (20)</td>
</tr>
<tr>
<td>Recovery</td>
<td>Non-Markovian</td>
<td>Delay: fixed as 14</td>
<td>(21)</td>
</tr>
</tbody>
</table>
By applying the estimated parameters and distributions, simulations were conducted using the modified Gillespie algorithm with 10,000 simulation runs for each scenario setting (22,23). To reflect the actual events in Korea, the superspreader was designated to stay in PMH for 3 d (May 15–17, 2015). The initial numbers of susceptible HCWs, patients, and visitors were 241, 263, and 389, respectively (14). The values in Table 1 represent the population distribution for the infectious period (20). However, due to factors such as isolation measures and ward closures in the hospital, the infectious period was adjusted by 25% (75% reduction) for use in the baseline scenario simulation (14,21). This adjustment ensured that the mean number of confirmed cases from the simulations matched the actual number of cases. The results for the adjustment ratio are introduced in the following section.

Since the onset of the Coronavirus disease pandemic, the wearing of masks in health facilities has been legally enforced until April 2024 (24). To quantify the effect of mask-wearing on preventing nosocomial spread, we conducted an additional scenario-based analysis. Parameters that had a significant impact on nosocomial spread were considered, and the following three scenarios were analysed:

✓ Analysis of the infectious period of hosts in the hospital: We adjusted the infectious period of the total population from 0% to 95% in our baseline scenario, which had already reduced it by 75%. The infectious period of the superspreader was fixed at 3 d.

✓ Analysis of adjusting the infectious period of a superspreader: We considered the infectious period for non-superspreaders as the baseline and adjusted the infectious period of the superspreader from 1 to 5 d.

✓ Analysis of mask-wearing interventions: We evaluated the impact of the mask mandate in hospitals. The effect varied according to the type of mask (76% for N95 and 30% for medical/surgical masks) and the level of enforcement (full effect for mandatory, half for recommended) (25). The preventive effect of wearing a mask is reflected in a reduction in the force of infection in susceptible hosts.
Results

Estimation of transmission rates

The parameters sampled using the Metropolis-Hastings algorithm are represented in the form of box-whisker plots in Figure 3. Figures 3 (A) and (B) show the transmission rates of non-superspreaders and superspreaders, respectively. The mean values of sampled transmission rates $\beta_{PH}$, $\beta_{PP}$, $\beta_{PV}$, $\beta_{VH}$, and $\beta_{VV}$ were 0.04, 0.61, 0.01, 0.17, 0.01, and 0.05 (95% confidence interval (CI) [0.00, 0.12], [0.42, 0.84], [0.00, 0.05], [0.01, 0.49], and [0.00, 0.16]), respectively, among which the transmission rate between patients ($\beta_{PP}$) was estimated to be the highest. The mean superspreader-induced transmission rates $\beta_{\Omega H}$, $\beta_{\Omega P}$, and $\beta_{\Omega V}$ were estimated to be 4.27, 15.04, and 10.57 (95% CI [1.69, 7.97], [9.89, 21.19], and [7.09, 14.82]). The superspreader-induced transmission rate among patients was estimated to be the highest.

Figure 3. Distribution of sampled transmission rates in Pyeongtaek St. Mary’s Hospital during the 2015 MERS outbreak in Korea: (A) Transmission rates induced by non-superspreaders; (B) transmission rates induced by a superspreader.
Simulation of the baseline scenario

The simulation results for the baseline scenario are shown in Figure 4. Figure 4 (A) shows the cumulative number of confirmed cases over time. The grey area in the graph represents the 95% credible interval (CrI), and the dark curve indicates the mean of the simulation runs. Figure 4 (B) shows the distribution of confirmed cases. The mean number of confirmed cases was 36.12, which was close to the actual number of cases, and the 95% CrI was from 23 to 50. The state variables \( E, I, \) and \( Q \) are visualised in Figure 4 (C), and the prevalence (proportion of \( E, I, \) and \( Q \) in the total population) is represented in Figure 4 (D). Based on the mean, we observed that \( E, I, \) and \( Q \) reached 28.62, 4.73, and 31.42 (3.3, 8.1, and 18.2 d after primary case onset). The possible peak in the 95% CrI prevalence was 6%. 
Figure 4. Baseline simulation results of the number of confirmed cases: (A) Cumulative confirmed cases over time; (B) distribution of the confirmed cases from simulation runs; (C) number of exposed, infectious, and isolated hosts over time; and (D) prevalence (proportion of exposed, infectious, and isolated hosts) over time. In panels A, B, and C, solid curves indicate mean value and coloured areas indicate 95% CrI.
Scenario-based study

To conduct the baseline scenario simulation, we reduced the infectious period by 75% compared to the population distribution, which includes the suspected infectious period outside the hospital, such that the mean value of the simulation runs follows the actual number of cases. Figure 5 shows the mean and 95% CrI of the confirmed cases, indicated by reduction varying from 0% to 95%. The cyan dotted vertical line indicates the value used in the baseline scenario, and the magenta horizontal dotted line represents the actual number of confirmed cases. The mean (95% CrI) number of confirmed cases changed from 100.40 ([39, 140]) to 30.70 ([22, 47]) as the reduction factor varies from 0% to 95%.

Figure 5. Number of confirmed cases varying with change in the infectious period reduction. Dark curve indicates simulation mean and red area covers 95% CrI.
In the baseline scenario, the infectious period of the superspreader was fixed at 3 d. To observe the impact on the scale of the outbreak when this period varies between 1 and 5 d, Figure 6 shows the distribution of confirmed cases according to the infectious period of the superspreader. The number of confirmed cases ranged from a minimum mean of 13.01 (95% CrI [6, 23]) to a maximum of 57.84 (95% CrI [41, 76]).

Figure 6. Distribution of the number of confirmed cases of the model simulation runs for the varying infectious period of a superspreader (1 to 5 d). Note that the distribution of baseline scenario simulation runs (3 d of infectious period) is not included in this figure.
Preventive effect of mask mandates

The results of the simulations considering the mask mandates are visualised in Figure 7. Figure 7 (A) indicates the type of mask worn by HCWs, patients, and visitors for each detailed scenario. Figure 7 (B) presents the simulation results considering mask mandates for each detailed scenario as an odds ratio of the number of confirmed cases compared with the baseline scenario, where mask mandates are not applied. The mean effect varied depending on the type of mask (N95 and medical or surgical masks), at 76% and 30%, respectively (25).

The highest reduction in the number of confirmed cases by mean 77% was achieved in the case of mandatory wearing of N95 or equivalent masks for everyone in the hospital. The lowest effect was achieved when recommending medical or surgical masks to everyone in the hospital, resulting in a mean reduction of 17% in the number of confirmed cases.

Figure 7. Simulation results of scenarios considering different mask mandates: (A) Description of scenario set-up; (B) odds ratio of confirmed cases. Recommendation level of intervention indicates that preventive effect of mask-wearing is reduced by half.
Discussions

This study encapsulates the process of estimating transmission rates using information collected from epidemiological investigations, specifically the suspected duration of the infectious period and the disease exposure time of individuals. By using the Metropolis-Hastings algorithm for parameter estimation, we were able to present the transmission rates as distributions, illustrating the uncertainty in the transmission rate of the underlying data.

During the MERS outbreak in 2015, the transmission rate in PMH was estimated to be the highest among patients, which was attributed to the spread originating from the inpatient ward. Similarly, the transmission rate induced by the superspreader was highest in the order of patients, visitors, and HCWs. The estimated transmission rate directly shows the risk of the spread of infectious diseases in hospitals. Excluding the superspreader, the mean transmission rate of 0.61 can be interpreted as the possibility of an additional 0.61 infections occurring per day by a single infectious patient.

The results of the baseline scenario simulation primarily show uncertainty in the scale of the outbreak, indicating that there could have been a minimum of 23 and up to 50 confirmed cases (Figure 3). Additionally, Figures 4 (C) and (D) show the point at which the prevalence within the hospital was expected to be greatest, which is 8 d after the onset in the baseline scenario. This indicates the time when the most infectious hosts were present. Paradoxically, this necessitates recognising the outbreak and tracking the infected individuals before the date when the spread within the hospital is most likely to occur.

Additional scenario-based analyses demonstrated the importance and potential of efforts to prevent the spread within the hospital. If the entire duration from symptom onset to isolation of infected individuals occurred within the hospital, meaning the reduction rate of the transmission period is 0% (Figure 5), there could have been more than 100 infected individuals in PMH alone. Moreover, if the period in which a single superspreader stayed in the hospital was extended to 5 d (an increase of 2 d compared to reality; Figure 6), there could have been 61% more cases.

Since the onset of the COVID-19 pandemic, mask mandates have been strongly implemented in Korea, and as of March 2024, it is mandatory to wear masks in hospitals. The results of this study show that maintaining such interventions has a significant effect
on preventing the spread of infectious diseases within hospitals. The study also provides appropriate mask intervention strategies depending on the level of spread prevention goals. For example, to maintain the prevention of infectious disease spread in hospitals at more than 50%, it is essential to take protective measures for HCWs and patients (Figure 7).

The overall methodology of this study can be applied to other outbreak situations, where it will yield different results due to spatial characteristics when applied to different countries, hospitals, or regions. This will help in ensuring preparedness and establishing intervention policies that can be tailored to the specific conditions in future outbreak situations.

The limitations of this study are as follows. (1) More detailed spaces, such as wards and rooms, within the hospital were not considered. (2) The effect of non-pharmaceutical interventions was considered to be average. (3) People visiting and leaving the hospital were not considered. (4) It was assumed that there was no transmission by HCWs. (5) We assumed a frequency-dependent transmission within hospitals. A detailed analysis considering the spatial arrangement within the hospital, patient tracking, isolation, testing, infection by HCWs, and different transmission types will be discussed in future work.

Conflict of interest statement

The authors declare no conflicts of interest.

Funding statement

This research was supported by the Government-wide R&D Fund Project for Infectious Disease Research (GFID), Republic of Korea (grant No. HG23C1629). This paper is supported by the Korea National Research Foundation (NRF) grant funded by the Korean government (MEST) (NRF-2021R1A2C100448711).
References


24. Starting today, you don’t have to wear a mask in the hospital... COVID-19 crisis level downgraded to ‘concern’, https://www.korea.kr/news/policyFocusView.do?newsId=148928706&bbsKey=&pageIndex=1 &pkgId=49500742&pkgSubId=&exceptNewsId=&cardYn=&catId=&catelDs=&sectId=&sectIds=&nCatelDs=&nSectIds=&dataValue=&srchWord=&repCode=&repText=; 2024 [accessed 3 June 2024].