A DOMAIN-KNOWLEDGE MODELING OF HOSPITAL-ACQUIRED INFECTION RISK IN HEALTHCARE PERSONNEL FROM RETROSPECTIVE OBSERVATIONAL DATA: A CASE STUDY FOR COVID-19

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Phat Kim Huynh

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Industrial Engineering and Management

May 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

A DOMAIN-KNOWLEDGE MODELING OF HOSPITAL-ACQUIRED INFECTION RISK IN HEALTHCARE PERSONNEL FROM RETROSPECTIVE OBSERVATIONAL DATA: A CASE STUDY FOR COVID-19

By

Phat Kim Huynh

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Trung (Tim) Quoc Le
Chair

Dr. Nita Yodo

Dr. Gang Shen

Approved:

05/20/2022
Date

Dr. David Grewell
Department Chair

ABSTRACT

Healthcare personnel (HCP) is facing a consistent risk of viral infections. We proposed a domain-knowledge-driven infection risk model to quantify the individual HCP and the population-level risks. For individual-level risk estimation, a time-variant model was proposed to capture the disease transmission dynamics. At the population-level, the infection risk was estimated using a Bayesian network model constructed from three feature sets. For model validation, we investigated the case study of the Coronavirus disease. The variance-based sensitivity analysis indicated that the uncertainty in the estimated risk was attributed to two variables: the number of close contacts and the viral transmission probability. We further validated the individual risk model by considering six occupations in the U.S. O*Net database. For the population-level risk model validation, the infection risk in Texas and California was estimated. The accurate estimation of infection risk will significantly enhance the PPE allocation, safety plans for HCP, and hospital staffing strategies.

ACKNOWLEDGEMENTS

I would like to give special thanks to the following people — MD. Arveity R. Setty, Dr. Om P. Yadav, Mr. Quan M. Tran, Dr. Nita Yodo, and Dr. Trung Q. Le, who have contributed to the research conceptualization, study design, and final manuscript revision. Their contributions are elaborated as following. MD. Arveity R. Setty defined the problems in the healthcare settings and verified the analytical methods and model formulation. Dr. Om P. Yadav proposed the sensitivity analysis method and provided the critical feedback about the model formalism and the data analysis methods. Mr. Quan M. Tran revised the thesis methodologies and developed the model formalism. Dr. Nita Yodo revised the methodologies and the overall logical flow of the thesis. Dr. Trung Q. Le contributed to the study design, the problem formulation, and the final manuscript revision. This work was supported by National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Number U54GM128729 and National Science Foundation under Award Number 2119691.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Nosocomial infections (i.e., hospital-acquired infections) of communicable viral diseases (CVDs) (e.g., influenza virus, hepatitis A virus, and rotavirus infections) have posed huge challenges to public health organizations. The U.S. Centers for Disease Control and Prevention (CDC) estimated roughly 1.7 million nosocomial infections contribute to 99,000 deaths each year [1]; additionally, other estimates indicate 2 million patients per year become infected, which requires the annual cost ranging from $4.5 billion to $11 billion [2]. Healthcare personnel (HCP) experience the highest risk [3-5] because of the direct or indirect contact with infected patients and virus-contaminated surfaces. Subsequently, these HCP may become widespread the virus to non-infectious patients, coworkers, and their family members. Although there has been an increasing number of hospital outbreaks of CVDs over the last decade, current containment and preventive measures in hospital settings usually overlook asymptomatic individuals and “super spreader” events [6, 7]. Hence a quantitative estimation of the infection risk in HCP is critical to mitigate and subsequently prevent nosocomial infections in hospitals. Furthermore, a precise measure of HCP infection risk is also important to address the epidemiological issues in hospital settings and provide information for personal protective equipment (PPE) allocation, safety plans for HCP, and staffing strategies.

Modeling nosocomial HCP infections in hospitals has been based on mathematical models to qualitatively capture the dynamics of CVDs and the effects of different control measures [8, 9]. One traditional model of disease spread is the compartmental SEIR (Susceptible-Exposed-Infected-Recovered) epidemiological model [10]. It divides a population into four different compartments or sub-groups (susceptible, exposed, infected, and recovered individuals) and employs deterministic ordinary differential equations to model the spread of a

CVD. In the literature, there are many variants of this model (e.g., SIS, SIRD, MSIR, and MSEIR model). These models consider the population as homogenous without individual interactions (e.g., patients and HCP); therefore, they fail to capture the individual contact process and the effects of individual risk and protective factors [11]. As mentioned, those models are often represented by a set of ordinary differential equations (i.e., deterministic systems), however they can also be augmented where stochastic components can be included to model more realistic scenarios. For demonstration, the compartmental SEIR model is depicted in **Figure 1.1**, where there is a substantially latency period over which people have been infected but not yet infectious. Therefore, these people can be included in the compartment E (for exposed), which models the number of individuals that are exposed to a viral disease but have not developed any symptoms. The SEIR model dynamics can be formulated as follows:

$$\frac{dS}{dt} = -\frac{\beta IS}{N} \tag{1.1}$$

$$\frac{dE}{dt} = \frac{\beta IS}{N} - \sigma E \tag{1.2}$$

$$\frac{dI}{dt} = \sigma E - \gamma I \tag{1.3}$$

$$\frac{dR}{dt} = \gamma I \tag{1.4}$$

where $\beta, \sigma,$ and γ are the model parameters, and N is the total population in the area of interest.

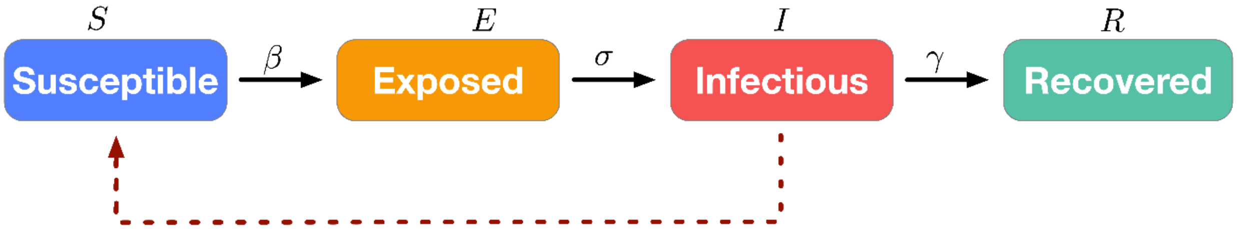

Figure 1.1. Illustration of the compartmental SEIR model, where $\beta, \sigma,$ and γ are the parameters that can be estimated from the reported diseased cases [12].

To overcome the limitations of the classic compartmental models, complex systems approaches using cellular automata (CA) theory have been proposed to model location-specific dynamics of susceptible populations and the probabilistic nature of disease transmission [13, 14]. The major drawback of CA models is their insufficiency in characterizing the spatial-temporal information of individuals' movements and interactions [15]. Agent-based modeling (ABM) was proposed to address the limitations of CA models by accounting for the movement of individual disease carriers and the contact network of people [16]. Although the ABM approach can capture the spread of a CVD in a spatial region (e.g., hospital) over time and estimate the risk of viral infection, it requires a large amount of information of individuals' movement and high computational cost. Moreover, individuals' movements are highly restricted in hospital settings, especially for patients who have positive test results for infectious diseases.

In addition, statistical models have been used as an alternative to mathematical models to quantify the effects of protective or risk factors on the time-variant infection risk of HCP. Statistical models capture the disease transmission dynamics within the hospital, HCP-related risk factors of infection, and other patients and HCP as sources of infection [17]. Two classes of statistical models, namely measure of association and statistical survival analysis, have been proposed to estimate HCP infection risk. The measure of association approaches quantifies the relationship between the exposed and diseased HCP groups by using the adjusted odds ratio (aOR), risk difference (RD), and relative risk (RR) as the risk measures [4, 18-21]. To capture the changes of HCP's characteristics and infection risk over time, survival analysis models are used to estimate the HCP infection risk and the expected duration of time until a viral infection occurs [22, 23]. Although time-dependent variables have been considered in the survival analysis models, the stochastic nature of epidemiological dynamics and individual interactions have not been investigated.

To overcome the above research gaps, in this thesis, we have proposed a probabilistic domain-knowledge model of the infection risk of CVDs for HCP. The proposed model was formulated for the infection risk estimation at both individual and population levels with respect to three modes of transmissions: 1) direct contact of susceptible HCP with other infectious individuals including patients and coworkers, 2) airborne viruses, and 3) contaminated equipment and surfaces. The individual-level risk model was built based on the population grouping in the SEIR model with the consideration of the time-varying confounders to capture the dynamical contagious disease transmission mechanism. At the population-level, three subsets of features, which will be introduced in **Sub-section 2.2**, were constructed and represented by a Bayesian network [24, 25], which was based on the work published in the *"Artificial Intelligence in Medicine"* journal: Huynh, Phat K., et al. "Probabilistic domain-knowledge modeling of disorder pathogenesis for dynamics forecasting of acute onset." *Artificial Intelligence in Medicine* 115 (2021): 102056. The main contributions of this thesis are 1) a novel probabilistic model characterizes the dynamics of the disease transmission in HCP over time and 2) a domain-knowledge risk analysis model that quantifies both the individual-level and population-level infection risk. The thesis has been submitted to the PLOS ONE journal as a journal article. The paper preprint is available at https://arxiv.org/ftp/arxiv/papers/2111/2111.05761.pdf. [26]. The remainder of the thesis is organized as follows: **Section 2** discusses the model formulation and validation; the sensitivity analysis results, and the COVID-19 case study are presented in **Section 3**, discussion and conclusion are provided in **Sections 4 and 5**.

2. METHODOLOGY

The proposed framework consists of two sub-models: (1) an individual-level infection risk model that quantifies the risk of infection of an HCP, and (2) a population-level model that estimates the infection risk under working conditions at a medical facility. The output from the first sub-model serves as an input for the estimation of the population infection risk in the second model. Other inputs, such as engineering control and administrative factors, were also considered in the estimation of population risk.

2.1. Individual infection risk model

The individual infection risk model aims to quantify the potential risk of infection associated with a healthcare worker subject to nosocomial infection, whose job functions require working in proximity of patients. The proposed individual-level infection risk model is formulated using the population grouping approach in the compartmental SEIR model [10], in which the population is divided into different compartments (i.e., Susceptible (S), Exposed (E), Infectious (I), or Recovered (R)). However, susceptible (S) and recovered individuals (R) cannot transmit the virus during the length of a hospital stay [10], hence we do not consider these compartments in our model. Moreover, we do not assume that the recovered patients confer immunity to reinfection when being released from isolation. HCP coworkers have also been shown to contribute significantly to virus spread within the healthcare setting if contracting a virus [22, 27]. To capture the virus transmission mechanism, the healthcare worker group (HW) is added to model the HCP-HCP transmission, and the infectious individuals are further classified into two sub-groups: the infection-confirmed group (IC) and the infection-suspected (IS) group. Infection-confirmed individuals are those who have lab-confirmed infections (*e.g.*, individuals have tested positive for COVID-19 using the polymerase chain reaction (PCR) test), and the infection-suspected group includes individuals who are suspected to have the virus infection because they developed symptoms but have never tested for the infectious disease. In total, four groups (E, IC, IS, HW) are considered to model the individual HCP infection risk. We denote the potential infection risk of the HCP j at location i (e.g., hospitals) over time from t_1 to t_2 as $PIR_{i,j}^{(t_1:t_2)}$, which is the cumulative risk of viral infection after contacting patients and contaminated surfaces. We denote $N_{E,j}^{(t_1:t_2)}, N_{IC,j}^{(t_1:t_2)}, N_{IS,j}^{(t_1:t_2)}$, and $N_{HW,j}^{(t_1:t_2)}$ as the number of exposed cases, infection-confirmed, infection-suspected, and colleagues that an HCP j has contacted with over the time $(t_1: t_2)$, which is denoted as $(\cdot)$ (e.g., $N_{E,j}^{(t_1:t_2)} = N_{E,j}^{(\cdot)}$).

An HCP j is assumed to have $CC_{X,k}^{(t_1:t_2)}$ independent close contacts with an individual k. Next, we denote $p_{X,k\to j}^{(\cdot)}$ as the probability of viral transmission from individual k to the HCP j, with $X \in \{E, IC, IS, HW\}$ being the compartment indicator of person k. Here, if the probability $p_{X,k\to j}^{(\cdot)}$ is constant, the viral transmission mechanism is modelled as a binomial process $Bin\left(CC_{X,k}^{(\cdot)}, p_{X,k\to j}^{(\cdot)}\right)$ [28], and there are $N_{E,j}^{(\cdot)} + N_{IC,j}^{(\cdot)} + N_{IS,j}^{(\cdot)} + N_{HW,j}^{(\cdot)}$ binomial processes in total. The sequence of contacts of HCP j ordered by time will be superscripted by person index $k(m)$ and compartment index $X(m)$ as follows:

$$\boldsymbol{C}^{(t_1:t_2)} = \left\{C_m^{k(m),X(m)} | k(m) = 1, \dots, N_{X(m),j}^{(\cdot)}\right\} \tag{2.1}$$

where $X(m) = \{E, IC, IS, HW\}$, m is the temporal order of close contacts from which the HCP j contracts the virus, $C_m^{k(m),X(m)} = 1$ if the HCP j contracts the virus at the m^{th} close contact, $C_m^{k(m),X(m)} = 0$ otherwise. As a result, the risk $PIR_{i,j}^{(\cdot)}$, is estimated as:

$$PIR_{i,j}^{(\cdot)} = \sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|} P\left(C_m^{k(m),X(m)} = 1, \boldsymbol{C}_{1:m-1}^{k(m),X(m)} = \boldsymbol{0}\right) \tag{2.2}$$

where $\left|\boldsymbol{C}^{(\cdot)}\right|$ is the total number of contacts and $\boldsymbol{C}_{1:m-1}^{k(m),X(m)}=\boldsymbol{0}$ means all previous $m-1$ contacts are the failed transmissions. Given the assumption of independent close contacts, **Eq. (2.2)** can be expressed as:

$$PIR_{i,j}^{(\cdot)}=\sum_{m=1}^{\left|\boldsymbol{C}^{(\cdot)}\right|}\left[\prod_{r=1}^{m-1}\left(1-p_{X(r),k(r)\rightarrow j}^{(\cdot)}\right)\right]p_{X(m),k(m)\rightarrow j}^{(\cdot)} \tag{2.3}$$

The expectation and variance of $PIR_{i,j}^{(\cdot)}$ are further investigated and presented in the in **Appendix B**. If we denote $TP_{-}^{j,k}$ and $TP_{+}^{j,k}$ as the patient admission time and the recovery time of an individual k with whom the HCP j has close contacts, the time interval $\left[TP_{-}^{j,k},TP_{+}^{j,k}\right]$ is the virus exposure period for the HCP j with the person k. Therefore, $p_{X(r),k(r)\rightarrow j}^{(t_1:t_2)}$ can be reduced to $p_{X(r),k(r)\rightarrow j}^{\left(\max\left\{t_1,TP_{-}^{j,k(r)}\right\}:\min\left\{t_2,TP_{+}^{j,k(r)}\right\}\right)}$. If $p_{X,k\rightarrow j}^{(\cdot)}$ is time-invariant, a logistic regression model is established to estimate the probability $p_{X(r),k(r)\rightarrow j}^{(\cdot)}$ as:

$$\log\left[\frac{p_{X(r),k(r)\rightarrow j}^{(\cdot)}}{1-p_{X(r),k(r)\rightarrow j}^{(\cdot)}}\right]=\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta} \tag{2.4}$$

$$p_{X(r),k(r)\rightarrow j}^{(\cdot)}=P\left(Y_{X(r),k(r)\rightarrow j}^{(\cdot)}=1\right)=\frac{\exp(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta})} \tag{2.5}$$

where $Y_{X(r),k(r)\rightarrow j}^{(\cdot)}$ is the indicator variable ($Y_{X(r),k(r)\rightarrow j}^{(\cdot)}=1$ means that HCP j has contracted the virus via the contact with person $k(r)$ and $Y_{X(r),k(r)\rightarrow j}^{(\cdot)}=0$ if HCP j has failed to contract the virus), $\boldsymbol{Z}$ is the covariate vector including the factors influencing the response and $\boldsymbol{\beta}$ is the coefficient vector

If $p_{X,k\rightarrow j}^{(\cdot)}$ varies over time, the constant $p_{X,k\rightarrow j}^{(\cdot)}$ assumption is relaxed by considering the cumulative distribution function that describes the probability of infection up to time t: $F(t)=$

$P(T \leq t) = 1 - \exp\left(-\int_0^t h(t)dt\right)$, in which T is the infection time and $h(t)$ is the hazard function. Hence, $PIR_{i,j}^{(\cdot)} = P(t_1 \leq T \leq t_2)$ is:

$$PIR_{i,j}^{(\cdot)} = 1 - e^{-\int_0^{t_2} h(t)dt} - e^{-\int_0^{t_1} h(t)dt} = \sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}\left(1 - e^{-\int_{\tau_m}^{\tau'_m} h_m(t)dt}\right) \quad (2.6)$$

$$h(t) = \begin{cases} 0 \; if \; t \notin [\tau_m, \tau'_m] \\ h_m(t) \; if \; t \in [\tau_m, \tau'_m] \end{cases}, \; \forall m = 1, \dots, |\boldsymbol{C}^{(\cdot)}| \quad (2.7)$$

where $[\tau_m, \tau'_m]$ is the time period of the m^{th} close contact with person $k(m)$, and $h_m(t)$ is the cumulative infection time distribution function for the m^{th} close contact**.** The probability $p_{X(r),k(r)\to j}^{(\cdot)}$ and $h_m(t)$ depend on various factors including HCP- dependent features, patient-dependent features, patient-HCP interactions, HCP-HCP interactions, and healthcare facilities' conditions.

2.2. Population risk indicator model

The population risk indicator quantifies the potential viral infection risk associated with a hospital/clinic over the time period $[t_1 : t_2]$. The population risk, annotated as $PIR_i^{(t_1:t_2)}$, is interpreted as the probability that an HCP contracts the disease under working conditions at place i given the information about the individual-level infection risk of all HCP at place i and the external factors. At this level, external factors from engineering and administrative controls within the hospital are considered. Those are the factors that affect the population-level infection risk apart from the individual-level risk. Representative examples of engineering controls are high-efficiency air, ventilation rates at the workplace, and infection isolation rooms for aerosol generating procedures. Administrative controls include formal HCP training regarding PPE availability level, training on risk factors and resources to promote personal hygiene. The $PIR_i^{(t_1:t_2)}$ is computed using logistic function as:

$$PIR_i^{(\cdot)} = \left\{1 + \exp\left[-\sum_j \frac{f\left(\boldsymbol{PIR}_{i,j=1,\dots n_{HCP}}^{(\cdot)}, \boldsymbol{F}\right)}{\tau}\right]\right\}^{-1} \tag{2.8}$$

where $\boldsymbol{PIR}_{i,j=1,\dots n_{HCP}}^{(\cdot)} = \left[PIR_{i,1}^{(\cdot)}, PIR_{i,2}^{(\cdot)}, \dots PIR_{i,n_{HCP}}^{(\cdot)}\right]^T$ is the vector of individual infection risk estimates of a total number of n_{HCP} HCP, τ is the scaling parameter, $\boldsymbol{F} = \{F_i\}$ is the vector of engineering control and administrative control factors. We denote $f(\cdot)$ as the abbreviated notation for the function of $PIR_{i,j}^{(\cdot)}$ and $\boldsymbol{F}$. When the working restriction policy is applied to a certain HCP, which forces him/her to be self-isolated at home, his/her individual risk will not be considered in that equation. The function $f(\cdot)$ can be simply formulated as a linear regression model such that:

$$f(\cdot) = \boldsymbol{\alpha}\boldsymbol{PIR}_{i,j=1,\dots n_{HCP}}^{(\cdot)} + w_1F_1 + \cdots + w_nF_n + b \tag{2.9}$$

where $\boldsymbol{\alpha}, w_i$, and b are the model parameters. Alternatively, the population risk $PIR_i^{(\cdot)}$ is estimated using a Bayesian network when we have access to the domain knowledge that describes the relationships between the control factors and the infection risk at the population level and individual level. Here, the Bayesian network model [25] is employed to incorporate the domain knowledge that influences the virus spread. The network is formulated based on three subsets of factors from the literature that affect the risk of infection including 1) individual-level factors, 2) engineering control factors, and 3) administrative control factors (see **Figure 2.1**). Individual-level factors include patient characteristics (e.g., time from exposure to symptom onset, clinical severity of patients), HCP-dependent factors (e.g., PPE sufficiency level, close contacts with patients, exposure level to infection, working hours per week), and intervention-related risks (e.g., endotracheal intubation, high flow nasal cannula (HFNC)). External factors include engineering control factors (e.g., ventilation rates, airborne infection isolation rooms) and administrative control factors (e.g., formal HCP training on PPE and disease risk factors, resources to promote personal hygiene). These factors are annotated as $\boldsymbol{ILF}$, $\boldsymbol{ECF}$, and $\boldsymbol{ACF}$, respectively. Hence, using the chain rule of the Bayesian network [29], the risk $PIR_i^{(\cdot)}$ is

$$PIR_i^{(\cdot)} = P\left(X_{PIR_i^{(\cdot)}} = 1 \middle| PIR_{i,j}^{(\cdot)}, \boldsymbol{ECF}, \boldsymbol{ACF}, \boldsymbol{ILF}\right) = \frac{P\left(X_{PIR_i^{(\cdot)}}, PIR_{i,j}^{(\cdot)}, \boldsymbol{ECF}, \boldsymbol{ACF}, \boldsymbol{ILF}\right)}{P(\boldsymbol{ECF})P(\boldsymbol{ACF})P(\boldsymbol{ILF}|\boldsymbol{ACF},\boldsymbol{ECF})\,P(PIR_{i,j}^{(\cdot)}|\boldsymbol{ILF})} \tag{2.10}$$

where $P(\cdot)$ is the probability function, and $X_{PIR_i^{(\cdot)}}$ is the indicator variable ($X_{PIR_{i,j}^{(\cdot)}} = 1$ indicates that an HCP contracts the disease and $X_{PIR_{i,j}^{(\cdot)}} = 0$ if they do not).

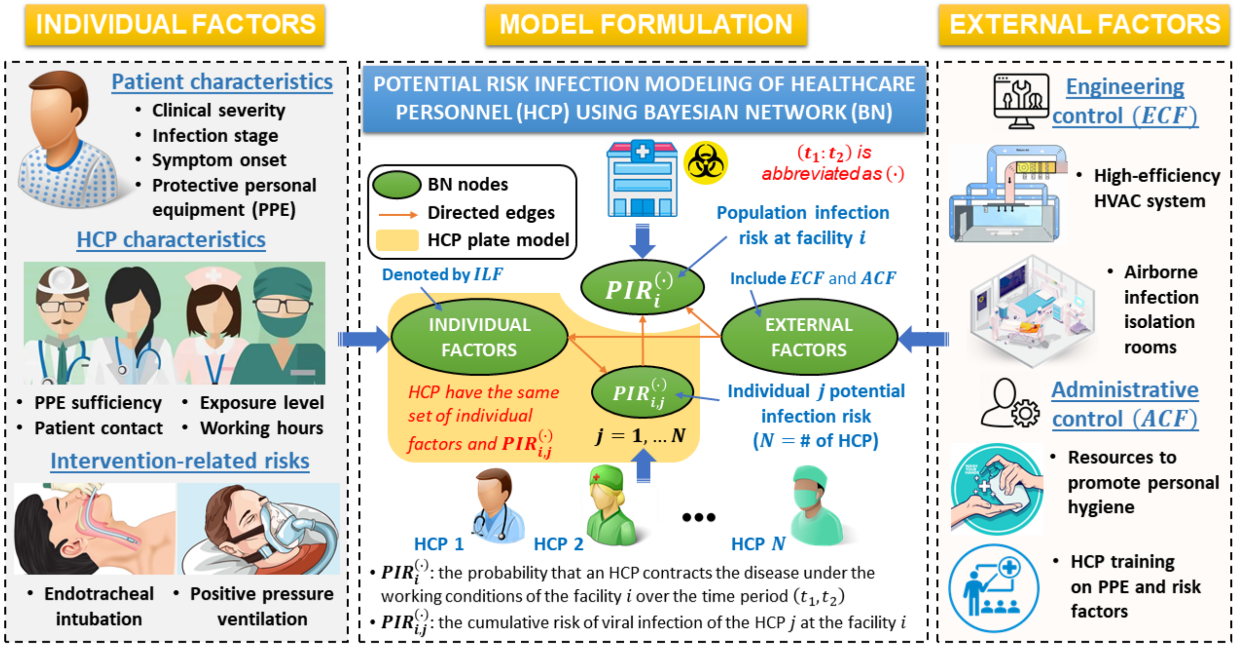
INDIVIDUAL FACTORS
MODEL FORMULATION
EXTERNAL FACTORS
Patient characteristics
• Clinical severity
• Infection stage
• Symptom onset
• Protective personal equipment (PPE)
HCP characteristics
• PPE sufficiency
• Patient contact
• Exposure level
• Working hours
Intervention-related risks
• Endotracheal intubation
• Positive pressure ventilation
POTENTIAL RISK INFECTION MODELING OF HEALTHCARE PERSONNEL (HCP) USING BAYESIAN NETWORK (BN)
BN nodes
Directed edges
HCP plate model
$(t_1 : t_2)$ is abbreviated as $(\cdot)$
Population infection risk at facility i
Denoted by ILF
Include ECF and ACF
$PIR_i^{(\cdot)}$
INDIVIDUAL FACTORS
EXTERNAL FACTORS
HCP have the same set of individual factors and $PIR_{i,j}^{(\cdot)}$
$PIR_{i,j}^{(\cdot)}$
$j = 1, \dots N$
Individual j potential infection risk (N = # of HCP)
HCP 1
HCP 2
HCP N
• $PIR_i^{(\cdot)}$: the probability that an HCP contracts the disease under the working conditions of the facility i over the time period (t_1, t_2)
• $PIR_{i,j}^{(\cdot)}$: the cumulative risk of viral infection of the HCP j at the facility i
Engineering control (ECF)
• High-efficiency HVAC system
• Airborne infection isolation rooms
Administrative control (ACF)
• Resources to promote personal hygiene
• HCP training on PPE and risk factors

Figure 2.1. Illustration of the contributions of the individual factors and external factors to the estimation of the infection risk of HCP in our model formulation. The infection risk at both individual level and population levels can be estimated based on a Bayesian network formulation which has 4 main nodes, namely the individual-level risk $PIR_{i,j}^{(\cdot)}$, the population-level risk $PIR_i^{(\cdot)}$, the individual-level factors, and the external factors. The individual-level factors ($\boldsymbol{ILF}$) include patient characteristics, HCP characteristics, and intervention-related risks, whereas the external factors consist of engineering control factors ($\boldsymbol{ECF}$) and administrative control factors ($\boldsymbol{ACF}$).

3. RESULTS AND COVID-19 CASE STUDY

3.1. Sensitivity analysis using simulated data

The variance-based sensitivity analysis was utilized to study the uncertainty of HCP's potential infection risk output caused by the variance of the input variables.

3.1.1. The measure of sensitivity of $\boldsymbol{PIR_{i,j}^{(\cdot)}}$ to $\boldsymbol{p_{X(m),k(m)\to j}^{(\cdot)}}$ and close contact sequence

The dependence of the infection risk on the probability of viral transmission and close contact sequence for an HCP was analyzed. The $PIR_{i,j}^{(\cdot)}$'s for different numbers of close contacts $\left|\boldsymbol{C}^{(\cdot)}\right|$ were estimated by **Eq. (2.4)**. For illustration, the results for $\left|\boldsymbol{C}^{(\cdot)}\right| = 2$ and $\left|\boldsymbol{C}^{(\cdot)}\right| = 3$ are shown in **Figure 3.1**.

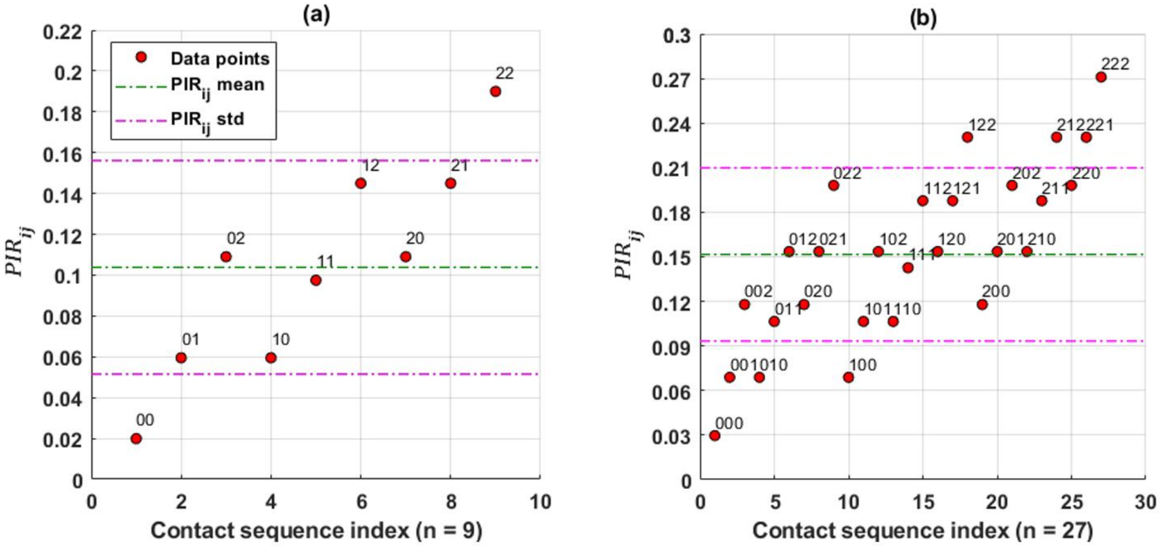

Figure 3.1. Sensitivity analysis of the impact of probability of viral transmission and the number of close contacts on $PIR_{i,j}^{(t_1:t_2)}$. We estimated values of $PIR_{i,j}^{(t_1:t_2)}$ for the synthesized data with three levels of $p_{X(m),k(m)\to j}^{(\cdot)}$: $P_{low} = 0.01, P_{medium} = 0.05, P_{high} = 0.1$. Panel (a): the estimated $PIR_{i,j}^{(t_1:t_2)}$ for $\left|\boldsymbol{C}^{(\cdot)}\right| = 2$, i.e., two close contacts; therefore, there are $n = 3^2 = 9$ possible contact sequences with different combinations of $p_{X(m),k(m)\to j}^{(\cdot)}$ levels, and those combinations are encoded in the form $X_1X_2 \ldots X_n$, where $X_1, X_2, \ldots, X_n \in \{0,1,2\}$, which corresponds to low, medium, and high levels of $p_{X(m),k(m)\to j}^{(\cdot)}$. The mean level of $PIR_{i,j}^{(t_1:t_2)}$ (green dash-dotted line) associated with its standard deviation indicated by purple dash-dotted lines are also plotted. Panel (b): the results for $\left|\boldsymbol{C}^{(\cdot)}\right| = 3$ with $n = 3^3 = 27$ possible contact sequences.

According to the results, the mean level ($\pm$ standard deviation) of $PIR_{i,j}^{(\cdot)}$ for $\left|\boldsymbol{C}^{(\cdot)}\right| = 2$ was 0.1038 ± 0.0523, which was lower than that for $\left|\boldsymbol{C}^{(\cdot)}\right| = 3$ at 0.1516 ± 0.0583. The mean value of the individual risk escalated together with the standard deviation values as the number of contacts increased. In addition, the estimated $PIR_{i,j}^{(\cdot)}$ was not influenced by the time order of the close contacts, e.g., the same $PIR_{i,j}^{(\cdot)} = 0.1065$ for three sequences: 011, 101, 110, where 0 and 1 are the encoded values for $P(Low)$ and $P(medium)$ respectively. The results are from the assumption of temporal independence between close contacts However, the risk would increase when the probability $p_{X(m),k(m)\to j}^{(\cdot)}$ for each contact raised to a higher value, hence the probabilities collectively contributed to the value of risk.

3.1.2. Response surface of the mean and variance of $\boldsymbol{PIR_{i,j}^{(t_1:t_2)}}$

The measure of sensitivity of potential infection risk $PIR_{i,j}^{(t_1:t_2)}$ of the HCP j at the place i over time $(t_1: t_2\ )$ was investigated. We denote the mean level and the variance of $PIR_{i,j}^{(t_1:t_2)}$ of all sequences given the number of close contacts $\left|\boldsymbol{C}^{(\cdot)}\right|$ as $E\left[PIR_{i,j}^{(\cdot)}\right]$ and $Var\left[PIR_{i,j}^{(\cdot)}\right]$, respectively. Next, we defined two levels of $p_{X(m),k(m)\to j}^{(\cdot)}$: $P_{low} \in (0{,}0.5]$ and $P_{high} = P_{low} + 0.3$, and derived the response surfaces of the $E\left[PIR_{i,j}^{(\cdot)}\right]$ and $Var\left[PIR_{i,j}^{(\cdot)}\right]$ with respect to two inputs P_{low} and $\left|\boldsymbol{C}^{(\cdot)}\right|$. As shown in **Figure 3.2**, the response surface of $E\left[PIR_{i,j}^{(\cdot)}\right]$ showed that a high probability of successful viral transmission $p_{X(m),k(m)\to j}^{(\cdot)}$ will result in an extremely high value of $E\left[PIR_{i,j}^{(\cdot)}\right]$, e.g., $E\left[PIR_{i,j}^{(\cdot)}\right] = 0.8336$ when $\left|\boldsymbol{C}^{(\cdot)}\right|$ is only 3, $P_{low} = 0.3$, and $P_{high} = 0.7$.

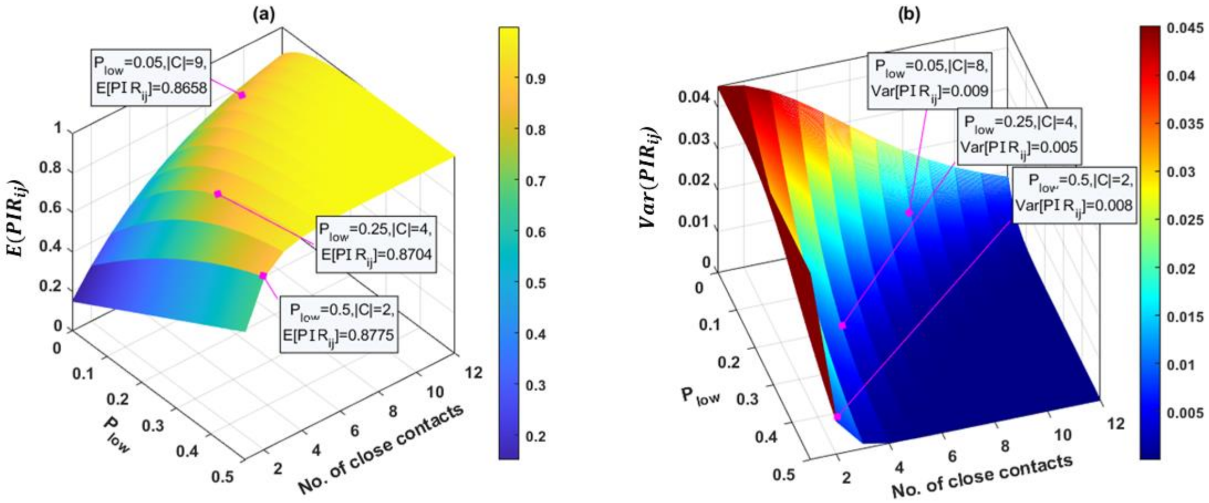

Figure 3.2. Response surfaces of $E\left[PIR_{i,j}^{(t_1:t_2)}\right]$ and $Var\left[PIR_{i,j}^{(t_1:t_2)}\right]$ with respect to two input variables: viral transmission probability and number of close contacts. (a): the response surface of $E\left[PIR_{i,j}^{(t_1:t_2)}\right]$ subject to the change of P_{low} and total number of close contacts $\left|\boldsymbol{C}^{(\cdot)}\right| \in [1,12]$. A data set was synthesized with two levels of $p_{X(m),k(m)\to j}^{(\cdot)}$: $P_{low} \in (0,0.5]$ and $P_{high} = P_{low} + 0.3$. where the expectation $E\left[PIR_{i,j}^{(t_1:t_2)}\right]$ is the mean level of $PIR_{i,j}^{(t_1:t_2)}$ of all possible contact sequences $\boldsymbol{C}^{(\cdot)}$, which are the combinations of P_{low} and P_{high} in the sequence of length $\left|\boldsymbol{C}^{(\cdot)}\right|$. Data tips at 3 values of P_{low}: $0.05, 0.2, 0.5$ were created to indicate the cut-off values of $\left|\boldsymbol{C}^{(\cdot)}\right|$ when $E\left[PIR_{i,j}^{(t_1:t_2)}\right]$ was significantly high. Similarly, (b) shows the response surface of $Var\left[PIR_{i,j}^{(t_1:t_2)}\right]$ of all possible sequences subject to the change of P_{low} and $\left|\boldsymbol{C}^{(\cdot)}\right|$. Three data tips at $P_{low} = \{0.05, 0.2, 0.5\}$ were included to show the threshold of $\left|\boldsymbol{C}^{(\cdot)}\right|$ at which $Var\left[PIR_{i,j}^{(t_1:t_2)}\right]$ was sufficiently low.

3.2. Model validation using COVID-19 case study

3.2.1. Case study description

Data sets of HCPs with COVID-19 were used to validate the proposed model. Access to these data sources can be provided per requests or via the cited references. The validation was performed on three main components: the viral transmission probability model, the individual-level infection risk model, and the population-level risk model. The HCP's occupational infection risk to COVID-19, interim guidance regarding risk assessment and universal PPE policy issued by the CDC [41], and the risk factors for severe acute respiratory syndrome coronavirus (SARS-CoV-2) transmission in hospital settings from previous studies were also included to develop the model for the case study.

The major factors resulting in high risk for HCPs are 1) exposure to COVID-19 patients without using appropriate PPE, 2) involvement in aerosol-generating procedures and the interventions performed by physicians or nurses, and 3) contact with patients and colleagues during the incubation period. Many studies suggested that there is a significant association between PPE use and infection risk and that masks are the most consistent contributing measure to reduce the risk [30, 31]. A similar association was observed for other PPE, such as gowns, gloves, and eye protection. Other exposures and treatment practices (e.g., intubation involvement, patient care, or having contact with secretions) were found to link with increased infection risk for HCPs [32, 33]. Finally, given the implementation of a universal PPE policy, the high risk of infection among HCP also arises from contacting asymptomatic patients and colleagues who are in the early phase of viral infections [20].

Different regression models, including logistic regression, log-binomial, and Poisson, were used with the defined risk measures to estimate the COVID-19 infection risk among HCP groups [19-21, 34-41]. Statistical survival analysis models were also used to estimate the HCP's risk of contracting SARS COV-2 viruses and the expected duration of time until viral infection occurs. Shah et al. [23] modeled hospital admission of healthcare workers with COVID-19 using Cox regression and conditional logistic regression. Long Nguyen et al. [22] assessed the COVID-19 infection risk among healthcare workers in contrast to the general community by examining the effect of PPE on risk. They also used Cox proportional hazards model to calculate multivariate-adjusted hazard ratios (HRs) of a positive test. However, the major limitations of these models are: 1) the individual-specific characteristics, e.g., occupation [42], type of PPE used, experience level, and exposure duration to COVID-19 patients, are not considered [22, 23], and 2) the simple formalism of the models without time-varying stochastic transmissions oversimplifies the complex contagious mechanism of SARS COV-2.

3.2.2. Data description

Data collected from multiple sources, namely COVID-19 transmission databases, health surveys/questionaries, U.S. Department of Labor databases, and cross-sectional study of UK-based healthcare workers, are illustrated in **Table 3.1**.

Table 3.1. Sources of databases information including source, nation, updated time, and owner

Data source	Nation	Updated time	Owner
Characteristics of HCP with COVID-19 [43]	US	July 16th, 2020	U.S. CDC
COVID-19 transmission dynamics data [44]	Taiwan	Apr 2nd, 2020	Taiwan CDC
California COVID-19 Health Surveys [45]	US	Sep 31st, 2020	California COVID-19 Health Center
Texas Health Center COVID-19 Survey [46]	US	Oct 7th, 2020	Texas Health Center
O*Net database [47]	US	Nov 16th, 2020	U.S. Department of Labor
COVID-NET database [48]	US	Aug 28th, 2020	U.S. CDC
Texas COVID-19 Data [49]	US	Apr 29th, 2021	Texas Department of State Health Services
Cross-sectional observational study of UK-based HCP [50]	UK	May 25th, 2020	The authors

3.2.3. Model variable selection

Variables from recent findings of SARS-CoV-2 as introduced in **Sub-section 3.2.1**, were used to select the features. The validation was performed on three main components: the viral transmission probability model, the individual-level infection risk model, and the population-level risk model. Regarding the viral transmission probability model, we included the following covariates in the model: Age, $Cancer$, $Resp$, $Obes$, $Smoker$, $Allied\_prof$, $Dental\_staff$, $Doctor$, Pub_trans, $C\_contact$, AGP, PPE_train, $Lacked\_PPE$, $Cont\_wo\_PPE$, and Imp_PPE. These are significant factors suggested by the original cross-sectional study [50]. The description of these variables is summarized in **Appendix A**. To validate the individual-level infection risk model, the U.S. Department of Labor O*Net database was employed to quantify the risk score for healthcare-related occupations, where virus exposure time and duration and working environment were considered. For the population-level risk model, the PPE sufficiency level, regional infection risk and the hospitalization data of HCP were selected to estimate population-level infection risk in California and Texas medical centers [45, 46] and implement a surrogate method for model validation. The description of these variables is summarized in **Appendix A**.

3.2.4. Model validation of viral transmission probability estimation using multivariate logistic regression

To validate the logistic regression introduced in **Sub-section 2.1**, we considered different protective and risk factors for COVID-19 in the data set of UK-based healthcare workers [50] and modelled the association between these covariates and the COVID-19 infection status using multivariable logistic regression. The data set provides 6263 responses in which a composite outcome was present in 1,806 (29.4%) HCP, of whom 49 (0.8%) HCP were admitted to hospitals, 459 (7.5%) were tested positive for SARS-CoV-2, and 1,776 (28.9%) HCP were self-isolated. The covariates included in the model were reported in **Sub-section 3.2.3** The estimated coefficients with their standard errors (SEs) and their statistical significance indicated by p-value are shown in **Table 3.2**.

Table 3.2. Estimated coefficients and their statistical significance for the multivariate logistic regression model

Variables	**Description**	**Coefficient estimates**	**SE**	**p-value**
Intercept	Intercept term	-0.5953	0.1497	6.98e-05***[a]
Age	Age of HCP	-0.5953	0.1497	6.98e-05***
$Cancer$	HCP's comorbidities include cancer	-0.0120	0.0028	1.77e-05***
$Resp$	HCP's comorbidities include respiratory disease	0.5296	0.2407	0.0277*
$Obes$	HCP's comorbidities include obesity	0.2020	0.0947	0.0328*
$Smoker$	HCP is a current smoker or ex-smoker within one year	0.3055	0.0872	0.0004***
$Doctor$	HCP is a current smoker or ex-smoker within one year	-0.2490	0.1053	0.0180*
$Allied\_prof$	HCP is a dentist or a dental staff	0.1514	0.0662	0.0222*
$Dental\_staff$	HCP is a doctor	-0.2282	0.0852	0.0074**
Pub_trans	HCP uses public transport to travel to work	-0.7018	0.2113	0.0008***
$C\_contact$	Having regular clinical contact with suspected or confirmed COVID-19 patients	0.2728	0.0693	8.31e-05***
AGP	Having regular exposure to aerosol generating procedures (AGPs) performed in suspected or confirmed COVID-19 patients	0.2949	0.0724	4.63e-05***
PPE_train	Having sufficient training in PPE use before handling patients	-0.2201	0.0663	0.0009***
$Lacked\_PPE$	Lacked access to PPE items for clinical contact with suspected or confirmed COVID-19 patients	-0.1708	0.0666	0.0104*
$cont\_wo\_PPE$	Frequency of contacting without PPE (classified into never, rarely, sometimes, often, and always)	0.3261	0.0768	2.21e-05***
Imp_PPE	HCP has used improvised (customized) PPE	-0.2070	0.0865	0.0166*

[a] Significance codes: $p \approx 0$ '***', $p < 0.001$ '**', $p < 0.01$ '*', AIC: 7317.7

According the table, the most significant variables (p-value < 0.001) that influence the disease transmission probability are Age, $Obes$, $Allied\_prof$, $Dental\_staff$, Pub_trans,

$C\_contact$, AGP, $Lacked\_PPE$, and $cont\_wo\_PPE$. The model goodness-of-fit was further assessed by the Akaike information criterion (AIC) and 10-fold cross validation. The AIC value for the above model was 7317.70 and that for the null model was 7449.75. The 10-fold cross validation accuracy was calculated to be 78.23%, which showed that the performance on test data was relatively good.

3.2.5. Model validation of the individual-level infection risk

To validate to infection risk model at the individual level, six occupations were considered using the U.S. Department of Labor O*Net database [47]. We also introduced a new variable called occupational-specific risk score denoted as ORS to account for the differences in infection risk among different occupations. The score was computed as:

$$ORS = \frac{(CO+PP+EI)}{3\phi} \times \frac{N_{hours}}{\max\{N_{hours}\}} \tag{3.1}$$

where $\max\{N_{hours}\}$ is the maximum working hours per week of 6 occupations, and ϕ is the scaling parameter. The description of those variables $CO, PP, EI,$ and N_{hours} are summarized in **Appendix A**. Because of the limited longitudinal data, our strategy was to validate the individual infection risk model using hypothesized scenarios of different occupational settings. Particularly, we made four main assumptions: 1) the individual-risk is the same for every individual working under the same conditions (e.g., same occupation), 2) all patients are confirmed cases, i.e., there is only one compartment IC, 3) the probabilities of viral transmission from all patients are the same for each occupation, and 4) the probability of viral transmission estimate for confirmed infectious patients, denoted as $\hat{p}_{IC}^{(t_1:t_2)}$, is equal to $ORS/\max\{ORS\}$, where $\max\{ORS\}$ is the maximum ORS score among 6 occupations, which guarantees $0 \leq p_{IC}^{(t_1:t_2)} \leq 1$. This is the surrogate approach for approximating the transmission probability defined in **Eq. (2.7)** in the scenario of limited individual-level data. Consequently, **Eq. (2.4)** is reduced to:

$$PIR_{i,j}^{(t_1:t_2)} = \sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|} \left(1 - \hat{p}_{IC}^{(t_1:t_2)}\right)^{m-1} \hat{p}_{IC}^{(t_1:t_2)} \tag{3.2}$$

Lastly, the total number of contacts $\left|\boldsymbol{C}^{(\cdot)}\right|$ was fixed to be 5 and the value ϕ was set to 20. Next, the risk was estimated using **Eq. (3.2)**, and the results are summarized in **Table 3.3**.

Table 3.3. Estimated individual-level infection risk for six different occupational settings

Occupations	$\boldsymbol{ORS}$	$\hat{\boldsymbol{p}}_{\boldsymbol{IC}}^{(\boldsymbol{t_1}:\boldsymbol{t_2})}$	$\boldsymbol{PIR}_{\boldsymbol{i,j}}^{(\boldsymbol{t_1}:\boldsymbol{t_2})}$
Registered Nurses	95.67	0.05	0.2262
Personal Care Aides	48.54	0.0254	0.1206
Nursing Assistants	59.08	0.0309	0.1451
Medical Assistants	89	0.0465	0.2119
Licensed Nurses	52.94	0.0277	0.1309
Respiratory Therapists	64.47	0.0337	0.1575

The results of the individual-level model indicated a strong positive association between the estimated risk $PIR_{i,j}^{(\cdot)}$ and virus transmission probability $p_{IC}^{(\cdot)}$, in which the top three occupations that have the highest risk were registered nurses, medical assistants, and respiratory therapists. Their associated $PIR_{i,j}^{(\cdot)}$ values were 0.2262, 0.2119, and 0.1575 respectively, which were relatively high when $\left|\boldsymbol{C}^{(\cdot)}\right| = 5$.

3.2.6. Model validation of the population-level infection risk

The population-level infection risk was validated based on the total of confirmed COVID-19 cases of HCP reported to the CDC. The number of positive COVID-19 cases of HCP in the US up to April 9, 2020, is presented in **Figure 3.3**.

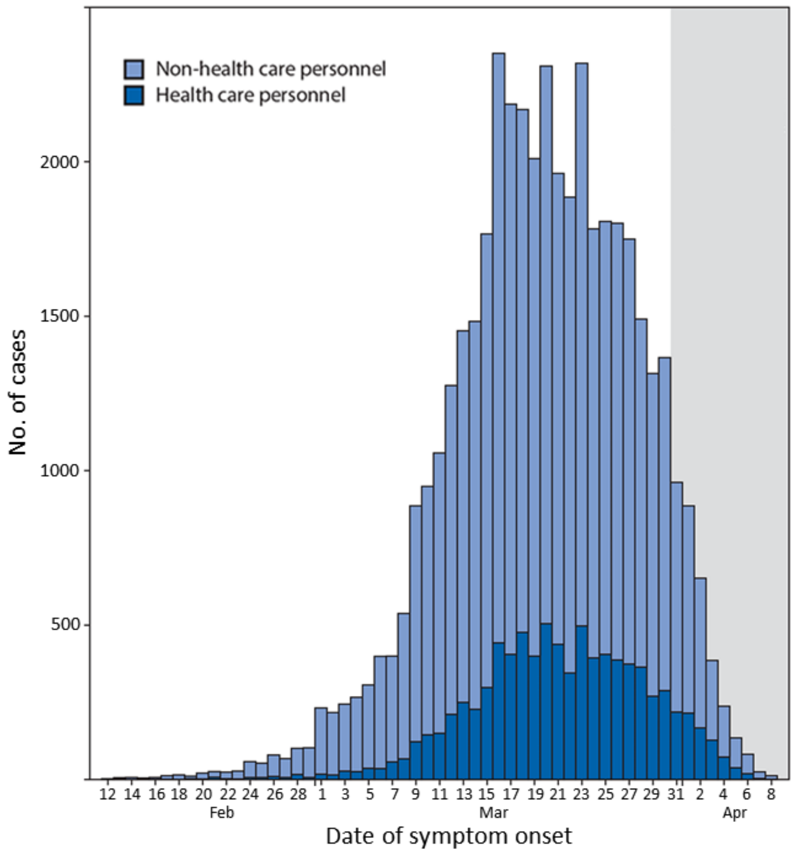

Figure 3.3. Daily number of laboratory-confirmed positive COVID-19 cases by date of symptom onset of health care personnel and non-health care personnel (N = 43968) in the US from February 12 to April 9, 2020 [43].

According to **Figure 3.3**, there was a strong association between the number of positive cases among non-HCP and the number of cases among HCP by date of symptom onset. In addition, the risk of infection among HCP was closely related to the total number of positive tests among HCP and the patient loads that HCP needed to handle. For population-level, we used the following selected features: SOH_{time}, CS, PPE_{SL}, ORS. The description of those is elaborated in **Appendix A**. Based on **Eq. (2.8)**, population-level risk estimation was reduced to a regressive equation with equal weights assigned to each variable as:

$$\widehat{PIR_\iota} = \frac{1}{4}\left(E_{SOH_{time}}\left[PIR_{i,j}^{(\cdot)}\right] + E_{CS}\left[PIR_{i,j}^{(\cdot)}\right] + E_{PPE_{SL}}\left[PIR_{i,j}^{(\cdot)}\right] + E_{ORS}\left[PIR_{i,j}^{(\cdot)}\right]\right) \qquad (3.3)$$

where $E_X\left[PIR_i^{(\cdot)}\right]$ is the expected value of $PIR_i^{(\cdot)}$ over the distribution of the variable X and $Value(X)$ is the value set of X, $E_X\left[PIR_i^{(\cdot)}\right]$ is estimated as:

$$E_X\left[PIR_i^{(\cdot)}\right] = \sum_{x \in Value(X)} P(X = x) E\left[PIR_i^{(\cdot)} \middle| X = x\right] \tag{3.4}$$

The population-level infection risk model was validated using the COVID-19 data from health centers in Texas, California and other relevant sources as presented in **Table 3.1**. The accessible HCP COVID-19 data of Texas and California were PPE sufficiency level, the total number of hospitalizations, and the percentage of ICU beds available. So, we assumed the distributions and the expected value of $PER_i^{(\cdot)}$ over the other variables to be the same for both states. The expected values of $PER_i^{(\cdot)}$ was computed using **Eq. (3.4)** (see **Table 3.4**).

Table 3.4. Estimated value and distribution of the selected features used in two case studies to estimate the infection risk in Texas and California

Features	**Texas**	**California**
Time from symptom onset to hospitalization	The distributions of SOH_{time} and CS are estimated from [44, 51]. $P(SOH_{time} < 0) = 0.34, P(SOH_{time} \in [0,3]) = 0.21$ $P(SOH_{time} \in [4,5]) = 0.05,$ $P(SOH_{time} \in [6,7]) = 0.02,$ $P(SOH_{time} \in [8,9]) = 0.16, P(SOH_{time} > 9) = 0.21$ $E_{SOH_{time}}\left[PIR_i^{(\cdot)}\right] = 5.99 \times 10^{-3}$	
Clinical severity of patients	$P(CS =$ "Severe pneumonia"$) = 0.01$, $P(CS =$ "ARDS/Sepsis"$) = 0.01$ $P(CS =$ "Asymptomatic"$) = 0.08$, $E_{CS}\left[PIR_i^{(\cdot)}\right] = 3.89 \times 10^{-3}$	
PPE sufficiency level	PPE sufficiency levels were averaged to estimate $E_{PPE_{SL}}\left[PIR_{i,j}^{(\cdot)}\right]$ to be 0.0065	$E_{PPE_{SL}}\left[PIR_{i,j}^{(\cdot)}\right]$ was estimated to be 0.744
ORS	$E_{ORS}\left[PIR_i^{(\cdot)}\right]$ was estimated to be the average over of $PIR_{i,j}^{(t_1:t_2)}$ over all occupations at 0.0173	
Estimated $\widehat{PIR}_i$	$\widehat{PIR}_{Texas} = 0.0084$	$\widehat{PIR}_{California} = 0.0132$

In **Table 3.4**, $E_{PPE_{SL}}\left[PIR_i^{(\cdot)}\right]$ was estimated using the PPE lacking information in health centers in Texas and HCP surveys in California. The value of $E_{ORS}\left[PIR_i^{(\cdot)}\right]$ was estimated by averaging the values of $PIR_{i,j}^{(\cdot)}$ over all occupations. The estimated $\widehat{PIR}_i$ values for Texas and California were 0.0084 and 0.0132, respectively.

4. DISCUSSION

Hospital-acquired infections of communicable viral diseases are posing a challenge to healthcare workers globally. HCP is facing a consistent risk of hospital-acquired infections, and subsequently higher rates of morbidity and mortality. Therefore, mitigating and preventing nosocomial infections in hospitals is an urgent and important task to lower the risk of contracting CVDs for HCP, guarantee adequate availability of PPE and develop well-informed strategies to protect health-care workers from contracting CVDs. In this thesis, we have developed a proposed probabilistic model characterizes the dynamics of the disease transmission in HCP over time, in which the domain-knowledge risk analysis framework can quantify both the individual-level and population-level infection risk. We validated the model at both levels using two main approaches, namely the variance-based sensitivity analysis using the simulated data and the COVID-19 case study. The sensitivity analysis indicated that the uncertainty in the HCP infection risk is attributed to 2 variables: the number of close contacts and the viral transmission probability. The COVID-19 case study showed that the occupations with the highest risk are registered nurses, medical assistants, and respiratory therapists. In addition, the results indicated the significant risk and protective factors of the COVID-19 transmission risk of HCP.

In our sensitivity analysis, we focused only on two key variables, namely viral transmission probability and the number of close contacts between HCP and patients. Specifically, the sensitivity of the infection risk to those input variables was measured by the amount of variance caused by changing the inputs. We divided our analysis into two parts: 1) the measure of sensitivity of $PIR_{i,j}^{(\cdot)}$ to $p_{X(m),k(m)\to j}^{(\cdot)}$ and close contact sequence, and 2) response surface of the mean and variance of $PIR_{i,j}^{(t_1:t_2)}$ to $\left|C^{(\cdot)}\right|$ and $p_{X(m),k(m)\to j}^{(\cdot)}$. The results of the sensitivity analysis revealed that the output $PIR_{i,j}^{(\cdot)}$ will be significantly increased when the viral transmission probability $p_{X(m),k(m)\to j}^{(\cdot)}$ and the number of contacts increases. In addition, the results in the second part indicated that $E\left[PIR_{i,j}^{(\cdot)}\right]$ quickly converged to one as $\left|\boldsymbol{C}^{(\cdot)}\right| \to \infty$, and the convergence rate was faster if P_{low} took higher values. Based on the response surface of $Var\left[PIR_{i,j}^{(\cdot)}\right]$, higher values of P_{low} and $\left|\boldsymbol{C}^{(\cdot)}\right|$ will lead to a lower value of $Var\left[PIR_{i,j}^{(\cdot)}\right]$; however, the effect of P_{low} is more significant than that of $\left|\boldsymbol{C}^{(\cdot)}\right|$. The value of $Var\left[PIR_{i,j}^{(\cdot)}\right] \to 0$ as $\left|\boldsymbol{C}^{(\cdot)}\right| \to \infty$ and dropped to nearly 0 after only four close contacts when $P_{low} = 0.5$.

After performing the sensitivity analysis, the logistic regression for estimating viral transmission probability $\hat{p}_{X,r\to j}^{(t_1:t_2)}$ was validated using the cross-sectional observational study of UK-based healthcare workers. Based on the coefficient estimates of the variables in the built multivariate logistic regression model, Age, $Smoker$, $Allied\_prof$, $Dental\_staff$, AGP, PPE_train, Imp_PPE were the protective factors, whereas the risk factors were $Cancer$, $Resp$, $Obes$, $Doctor$, $Lacked\_PPE$, $cont\_wo\_PPE$, Pub_trans, $C\_contact$. Surprisingly, advanced age, being a smoker or ex-smoker within one year, and having regular exposure to aerosol-generating procedures performed on COVID-19 patients decreased the infection risk. This result seems counter-intuitive at first, but they are confounders because it was shown that HCP working directly with suspected or confirmed COVID-19 patients tended to be more cautious and self-aware in clinical environments[52]. Therefore, they had sufficient self-protection and took containment measures; however, healthcare workers in non-communicable viral disease departments, who were potentially exposed to contagious viruses, did not have sufficient training on how to use PPE and deal with infectious diseases and lack of access to PPE and isolation equipment [53]. However, the model has several limitations. First, because we did not have access to information on HCP contact with patients and coworkers, we assumed the estimated viral transmission probability as a measure averaged over all individuals. Second, the data were gathered using surveys and questionnaires, which are subject to selection and recall bias. Third, the use of a composite outcome (including HCP with COVID-19 symptoms, HCP being exposed to risk factors, and lab-confirmed HCP infections) may have resulted in overestimation or underestimation of the infection risk.

We validated the individual-level infection risk model, implemented the model using the two-parameter regressive equation, and estimated the individual risk for six occupations. The results highly depend on the pre-defined parameters, which can be estimated in healthcare settings when data are available. It was shown that healthcare workers and nurses are frequently in close contact with COVID-19 patients, which therefore increases the risk for acquiring SARS-CoV-2 virus [54]. Because HCP can acquire infection through various pathways apart from direct patient care, such as exposure to colleagues, family members, or people in the community, the time-varying risk estimation in the model can provide informed decisions for screening HCP for COVID-19 before workplace entry. The individual risk model can be improved and more specific to better model the transmission dynamics, e.g., a model that incorporates the quantification of indoor airborne infection risks using a probabilistic framework [55]. In addition, we do not assume that the recovered patients confer immunity to reinfection when being released from isolation This statement can be further clarified that even the patients are fully recovered after getting contracted with COVID-19 (or any communicable viral diseases in general) and being released back to the population, they are still under the risk of reinfection with the same disease strain or other strains. However, reinfection with the same strain is very rare. Hence, if the HCP were recovered from the disease, they might get infected again; however, they still confer some degree of immunity from subsequent infection. Therefore, we can consider adding a new group of patients in our model called “reinfected patients”. Moreover, the same idea can be applied for the vaccinated population, in which people have vaccine-induced immunity.

For model validation at the population level, we considered two case studies to estimate the risk of infection of HCP in Texas and California states. Both states have a high number of lab-confirmed SARS-CoV-2 patients. The average number of hospitalizations in Texas and California were 16843 cases/day and 4219 cases/day, respectively. However, the infection risk in Texas was 0.0084 which was lower than the risk in California (0.0132). This was mainly due to the difference in patient load for each HCP per day and the two states’ PPE sufficiency level. From **Table 3.4**, the average PPE sufficiency level in California was only 0.744 as opposed to 0.9355 in Texas, and the average percentage of ICU beds available per 100,100 people in Texas was significantly higher than that in California, which implies heavier patient loads in California. The model also made some important assumptions: 1) close contacts with COVID-19 patients are independent and there is no viral transmission among HCP, and 2) protective/risk factors are well-defined and sufficient to estimate the risk of infection.

5. CONCLUSION AND FUTURE WORK

The thesis proposed a time-variant infection risk analysis model to characterize the dynamic of the disease infection risk in HCP over time and a domain-knowledge driven infection risk to quantify the complexities of HCP's risk of CVDs in healthcare settings. The infection risk analysis model for HCP was estimated at both individual and population levels. The individual-level risk model was built based on the population grouping concept of the well-established epidemiological SEIR model with the consideration of the time-varying confounders to capture the dynamical contagious disease transmission mechanism. At the population-level, three subsets of features were constructed and represented by a Bayesian network, from which the probability of viral transmission from patients to HCP was estimated. To validate our methods, we have incorporated the data from multiple data sources from the US, the UK, and Taiwan for the COVID-19 case study, which contains the information about potential factors that affect COVID-19 transmission mechanism; and the domain knowledge of similar contagious diseases such as SARS or MERS from the relevant studies to estimate the risk of COVID-19 infection of HCP. For individual-level risk estimation, the model was founded on the SEIR compartmental model and developed for the occupational-specific and individualized infection risk model. As a result, the model can accurately capture the infection risk varying over time under the control of those individual time-varying confounders, and it is also able to account for the intrinsic stochastic transmission mechanisms. At the population level, the Bayesian network formalism can accommodate the limited data scenario, and it can update the parameters when more data are available. The results from two case studies are interpretable at the population level, which showed infection risk in California is higher than in Texas because of the heavier patient loadings and shortage of PPE. The major limitations of the CDC's interim guideline for risk assessment, which is inadequate in quantifying the risk of infection in an individualized HCP, have been addressed by our model. The model would significantly endorse the PPE allocation and safety plans for HCP and enhance the crisis-level staffing strategies in facilities with the staffing shortages. Longitudinal experimental designs are required to collect more COVID-19 data among HCP to validate the proposed model properly. Future work would involve: 1) model assumption validation when more data are available and sufficient, 2) model modification and reformulation if the assumptions are violated (e.g., independence assumption and new vaccinated or "reinfected" population), and 3) validating the model with the other related case studies of communicable viral diseases.

REFERENCES

1. Klevens, R.M., et al., *Estimating health care-associated infections and deaths in US hospitals, 2002.* Public health reports, 2007. **122**(2): p. 160-166.
2. Hensley, B.J. and J.R. Monson, *Hospital-acquired infections.* Surgery (Oxford), 2015. **33**(11): p. 528-533.
3. Iversen, K., et al., *Risk of COVID-19 in health-care workers in Denmark: an observational cohort study.* The Lancet Infectious Diseases, 2020. **20**(12): p. 1401-1408.
4. Mutambudzi, M., et al., *Occupation and risk of severe COVID-19: prospective cohort study of 120 075 UK Biobank participants.* Occupational and Environmental Medicine, 2020.
5. Baker, M.G., T.K. Peckham, and N.S. Seixas, *Estimating the burden of United States workers exposed to infection or disease: a key factor in containing risk of COVID-19 infection.* PloS one, 2020. **15**(4): p. e0232452.
6. McDougal, A.N., et al., *Outbreak of coronavirus disease 2019 (COVID-19) among operating room staff of a tertiary referral center: An epidemiologic and environmental investigation.* Infection Control & Hospital Epidemiology, 2021: p. 1-7.
7. Khatib, A.N., et al., *Navigating the risks of flying during COVID-19: a review for safe air travel.* Journal of travel medicine, 2020. **27**(8): p. taaa212.
8. Cooper, B.S., *Confronting models with data.* Journal of Hospital Infection, 2007. **65**: p. 88-92.
9. Grundmann, H. and B. Hellriegel, *Mathematical modelling: a tool for hospital infection control.* The Lancet infectious diseases, 2006. **6**(1): p. 39-45.
10. Brauer, F., *Compartmental models in epidemiology*, in *Mathematical epidemiology*. 2008, Springer. p. 19-79.
11. Di Stefano, B., H. Fuks, and A.T. Lawniczak. *Object-oriented implementation of CA/LGCA modelling applied to the spread of epidemics*. in *2000 Canadian Conference on Electrical and Computer Engineering. Conference Proceedings. Navigating to a New Era (Cat. No. 00TH8492)*. 2000. IEEE.
12. Zou, D., et al., *Epidemic model guided machine learning for COVID-19 forecasts in the United States.* MedRxiv, 2020.
13. Sirakoulis, G.C., I. Karafyllidis, and A. Thanailakis, *A cellular automaton model for the effects of population movement and vaccination on epidemic propagation.* Ecological Modelling, 2000. **133**(3): p. 209-223.
14. Zhen, J. and L. Quan-Xing, *A cellular automata model of epidemics of a heterogeneous susceptibility.* Chinese Physics, 2006. **15**(6): p. 1248.
15. Casalicchio, E., E. Galli, and S. Tucci, *Agent-based modelling of interdependent critical infrastructures.* International Journal of System of Systems Engineering, 2010. **2**(1): p. 60-75.
16. Perez, L. and S. Dragicevic, *An agent-based approach for modeling dynamics of contagious disease spread.* International journal of health geographics, 2009. **8**(1): p. 1-17.
17. Voirin, N., et al., *A multiplicative hazard regression model to assess the risk of disease transmission at hospital during community epidemics.* BMC medical research methodology, 2011. **11**(1): p. 1-8.

18. Chu, D.K., et al., *Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis.* The lancet, 2020. **395**(10242): p. 1973-1987.
19. Wang, Q., et al., *Epidemiological characteristics of COVID-19 in medical staff members of neurosurgery departments in Hubei province: a multicentre descriptive study.* medRxiv, 2020.
20. Eyre, D.W., et al., *Differential occupational risks to healthcare workers from SARS-CoV-2 observed during a prospective observational study.* Elife, 2020. **9**: p. e60675.
21. Ki, H.K., et al., *Risk of transmission via medical employees and importance of routine infection-prevention policy in a nosocomial outbreak of Middle East respiratory syndrome (MERS): a descriptive analysis from a tertiary care hospital in South Korea.* BMC pulmonary medicine, 2019. **19**(1): p. 1-12.
22. Nguyen, L.H., et al., *Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study.* The Lancet Public Health, 2020. **5**(9): p. e475-e483.
23. Shah, A.S., et al., *Risk of hospital admission with coronavirus disease 2019 in healthcare workers and their households: nationwide linkage cohort study.* bmj, 2020. **371**.
24. Friedman, N., D. Geiger, and M. Goldszmidt, *Bayesian network classifiers.* Machine learning, 1997. **29**(2-3): p. 131-163.
25. Huynh, P.K., et al., *Probabilistic domain-knowledge modeling of disorder pathogenesis for dynamics forecasting of acute onset.* Artificial Intelligence in Medicine, 2021. **115**: p. 102056.
26. Huynh, P.K., et al., *A Probabilistic Domain-knowledge Framework for Nosocomial Infection Risk Estimation of Communicable Viral Diseases in Healthcare Personnel: A Case Study for COVID-19.* arXiv preprint arXiv:2111.05761, 2021.
27. Chou, R., et al., *Epidemiology of and risk factors for coronavirus infection in health care workers: a living rapid review.* Annals of internal medicine, 2020. **173**(2): p. 120-136.
28. Kallenberg, O., *Random measures, theory and applications*. Vol. 1. 2017: Springer.
29. Jensen, F.V., *Bayesian networks.* Wiley Interdisciplinary Reviews: Computational Statistics, 2009. **1**(3): p. 307-315.
30. Mo, Y., et al., *Transmission of community-and hospital-acquired SARS-CoV-2 in hospital settings in the UK: A cohort study.* PLoS medicine, 2021. **18**(10): p. e1003816.
31. Lan, F.-Y., et al., *COVID-19 symptoms predictive of healthcare workers' SARS-CoV-2 PCR results.* PloS one, 2020. **15**(6): p. e0235460.
32. Chen, Q., A. Allot, and Z. Lu, *Keep up with the latest coronavirus research.* Natur, 2020. **579**(7798): p. 193-193.
33. Raboud, J., et al., *Risk factors for SARS transmission from patients requiring intubation: a multicentre investigation in Toronto, Canada.* PLoS One, 2010. **5**(5): p. e10717.
34. Caputo, K.M., et al., *Intubation of SARS patients: infection and perspectives of healthcare workers.* Canadian Journal of Anesthesia, 2006. **53**(2): p. 122.
35. Chen, W.-Q., et al., *Which preventive measures might protect health care workers from SARS?* BMC Public Health, 2009. **9**(1): p. 1-8.
36. Le Dang Ha, S.A.B., et al., *Lack of SARS transmission among public hospital workers, Vietnam.* Emerging infectious diseases, 2004. **10**(2): p. 265.

37. Alraddadi, B.M., et al., *Risk factors for Middle East respiratory syndrome coronavirus infection among healthcare personnel.* Emerging infectious diseases, 2016. **22**(11): p. 1915.

38. Hall, A.J., et al., *Health care worker contact with MERS patient, Saudi Arabia.* Emerging infectious diseases, 2014. **20**(12): p. 2148.

39. Bai, Y., et al., *SARS-CoV-2 infection in health care workers: a retrospective analysis and a model study.* medRxiv, 2020.

40. Heinzerling, A., et al., *Transmission of COVID-19 to health care personnel during exposures to a hospitalized patient—Solano County, California, February 2020.* 2020.

41. Mutambudzi, M., et al., *Occupation and risk of severe COVID-19: prospective cohort study of 120 075 UK Biobank participants.* Occupational and Environmental Medicine, 2021. **78**(5): p. 307-314.

42. Lan, F.-Y., et al., *Work-related COVID-19 transmission in six Asian countries/areas: a follow-up study.* PloS one, 2020. **15**(5): p. e0233588.

43. Burrer, S.L., et al., *Characteristics of health care personnel with COVID-19—United States, February 12–April 9, 2020.* Morbidity and mortality weekly report, 2020. **69**(15): p. 477.

44. Cheng, H.-Y., et al., *Contact tracing assessment of COVID-19 transmission dynamics in Taiwan and risk at different exposure periods before and after symptom onset.* JAMA internal medicine, 2020.

45. *California Health Care Foundation, California COVID-19 Health Surveys: Data and Charts*. April 1, 2020; Available from: https://www.chcf.org/project/california-covid-19-health-surveys/#physician-survey.

46. *Health Resources & Services Adminstration, Texas Health Center COVID-19 Survey Summary Report*. Oct 7th, 2020; Available from: https://bphc.hrsa.gov/emergency-response/coronavirus-health-center-data/tx.

47. *O*Net database*. Nov 16th, 2020; Available from: https://www.onetonline.org/.

48. Garg, S., *Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 States, March 1–30, 2020.* MMWR. Morbidity and mortality weekly report, 2020. **69**.

49. *Texas COVID-19 Data*. Apr 29th, 2021; Available from: https://dshs.texas.gov/coronavirus/additionaldata.aspx.

50. Kua, J., et al., *healthcareCOVID: a national cross-sectional observational study identifying risk factors for developing suspected or confirmed COVID-19 in UK healthcare workers.* PeerJ, 2021. **9**: p. e10891.

51. COVID, T.C., *Characteristics of Health Care Personnel with COVID-19-United States, February 12-April 9, 2020.* 2020.

52. Du, Q., et al., *Nosocomial infection of COVID-19: A new challenge for healthcare professionals.* International Journal of Molecular Medicine, 2021. **47**(4): p. 1-1.

53. McMichael, T.M., et al., *Epidemiology of Covid-19 in a long-term care facility in King County, Washington.* New England Journal of Medicine, 2020. **382**(21): p. 2005-2011.

54. Hughes, M.M., et al., *Update: characteristics of health care personnel with COVID-19—United States, February 12–July 16, 2020.* Morbidity and Mortality Weekly Report, 2020. **69**(38): p. 1364.

55. Liao, C.M., C.F. Chang, and H.M. Liang, *A probabilistic transmission dynamic model to assess indoor airborne infection risks.* Risk Analysis: An International Journal, 2005. **25**(5): p. 1097-1107.
56. Chaloner, K.M. and G.T. Duncan, *Assessment of a beta prior distribution: PM elicitation.* Journal of the Royal Statistical Society: Series D (The Statistician), 1983. **32**(1-2): p. 174-180.

APPENDIX A. CHARACTERISTICS OF THE SELECTED FEATURES AND THEIR ASSOCIATED DATABASES

	Features	Values/Units	Notation	Characteristics	Data sources
	Time from symptom onset to hospitalization	Days	SOH_{time}	Secondary clinical attack rate is significantly high within the first 5 days from symptom onset	COVID-19 transmission dynamics data in Taiwan
	Clinical severity of patients	Discrete	CS	Be classified into: Asymptomatic, Mild illness, Mild pneumonia, Severe pneumonia, and ARDS/sepsis	
	PPE sufficiency level	%	PPE_{SL}	To assess the sufficiency level, we used the answer of nurses and physicians to the question: “Does your hospital have adequate PPE for clinicians to treat the patients you have right now?”. In addition, Texas Health Center COVID-19 Survey Summary Report provided the health centers with an adequate supply of (PPE) for the next week.	+ California COVID-19 Health Surveys: Data and Charts + Texas Health Center COVID-19 Survey Summary Report
	Contact with others	Score from 0 to 1	CO	The four physical job attributes help to determine the occupational-specific risk score: • **Contact with others**: How much does this job require the worker to be in contact with others to perform it? • **Physical proximity**: To what extent does this job require the worker to perform tasks in close physical proximity to others? • **Exposure to disease/infection**: How often does this job require exposure to disease or infection? • **Working hours per week**	U.S. Department of Labor O*Net database
	Physical proximity		PP		
	Exposure to disease/infection		EI		
	Working hours per week		N_{hours}		
Patient characteristics	Age	Continuous	Age	Age of HCP	Cross-sectional observational study of UK-based healthcare workers
	Having Cancer	Binary	$Cancer$	HCP’s comorbidities include cancer	
	Having respiratory disease	Binary	$Resp$	HCP’s comorbidities include respiratory disease	
	Having obesity	Binary	$Obes$	HCP’s comorbidities include obesity	
	Current or Ex-smoker within 1 year	Binary	$Smoker$	HCP is a current smoker or ex-smoker within one year	

Features		Values/Units	Notation	Characteristics	Data sources
Work details	Allied health professionals	Binary	$Allied\_prof$	HCP is a current smoker or ex-smoker within one year	Cross-sectional observational study of UK-based healthcare workers
	Dentists and dental staffs	Binary	$Dental\_staff$	HCP is a dentist or a dental staff	
	Doctors	Binary	$Doctor$	HCP is a doctor	
	Use public transport	Binary	Pub_trans	HCP uses public transport to travel to work	
Workplace exposure	Regular clinical contact	Discrete	$C\_contact$	Having regular clinical contact with suspected or confirmed COVID-19 patients	
	Regular exposure to AGPs	Discrete	AGP	Having regular exposure to aerosol generating procedures (AGPs) performed in suspected or confirmed COVID-19 patients	
PPE usage	Sufficient training in PPE use	Binary	PPE_train	Having sufficient training in PPE use before handling patients	
	Lacked access to PPE	Binary	$Lacked\_PPE$	Lacked access to PPE items for clinical contact with suspected or confirmed COVID-19 patients	
	Clinical contact without adequate PPE	Discrete	$Cont\_wo\_PPE$	Be classified into never, rarely, sometimes, often, always	
	Used improvised PPE	Binary	Imp_PPE	HCP has used improvised (customized) PPE	

APPENDIX B. EXPECTATION AND VARIANCE OF THE POTENTIAL INFECTION RISK

Here, we assume that all contraction trials are independent within and between infection groups. The expected potential infection risk is estimated as:

$$E\left[PIR_{i,j}^{(t_1:t_2)}\right] = E\left[\sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}\prod_{r=1}^{m-1}\left(1-p_{X(r),k(r)\to j}^{(\cdot)}\right)p_{X(m),k(m)\to j}^{(\cdot)}\right]$$

$$= \sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}\left[\prod_{r=1}^{m-1}\left(1-Ep_{X(r),k(r)\to j}^{(\cdot)}\right)Ep_{X(m),k(m)\to j}^{(\cdot)}\right] \quad \text{(B.1)}$$

Let's assume that the viral transmission from a person k to an HCP j probability $p_{X,k\to j}^{(\cdot)} \sim Beta(\alpha_{X,k},\beta_{X,k})$, where $X \in \{S,E,I,HW\}$, because it is a canonical conjugate prior to the binomial distribution and a common prior for expert elicitation when limited data are available [56]. When more data are accumulated, the likelihood term would dominate prior distribution. Using the properties of beta distribution gives:

$$E\left[PIR_{i,j}^{(t_1:t_2)}\right] = \sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}\prod_{r=1}^{m-1}\left(\frac{\beta_{X(r),k(r)}}{\alpha_{X(r),k(r)}+\beta_{X(r),k(r)}}\right)\frac{\alpha_{X(m),k(m)}}{\alpha_{X(m),k(m)}+\beta_{X(m),k(m)}} \quad \text{(B.2)}$$

The variance $Var\left[PIR_{i,j}^{(t_1:t_2)}\right]$ is:

$$Var\left[PIR_{i,j}^{(t_1:t_2)}\right] = Var\left[\sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}\prod_{r=1}^{m-1}\left(1-p_{X(r),k(r)\to j}^{(\cdot)}\right)p_{X(m),k(m)\to j}^{(\cdot)}\right] \quad \text{(B.3)}$$

Let's denote $X_r \coloneqq p_{X(r),k(r)\to j}^{(\cdot)}, Y \coloneqq p_{X(m),k(m)\to j}^{(\cdot)}$, and $Z_m \coloneqq \prod_{r=1}^{m-1}(1-X_r)Y$. We can assume $X_r, Y,$ and Z_m are i.i.d., which gives $Cov(Z_i,Z_j) = 0\ \forall i \neq j$, we have:

$$Var\left(\sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}Z_m\right) = \sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}Var(Z_m) + \sum_{i\neq j}Cov(Z_i,Z_j) = \sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}Var(Z_m) \text{(B.4)}$$

Therefore, to estimate $Var\left(\sum_{m=1}^{|\boldsymbol{C}^{(\cdot)}|}Z_m\right)$, we calculate the variance $Var(Z_m)$ where $m = 1,\dots,|\boldsymbol{C}^{(\cdot)}|$:

$$Var(Z_m) = Var[\prod_{r=1}^{m-1}(1 - X_r)Y] \tag{B.5}$$

Let's assume that $X_i, Y \perp X_j^n \; \forall i \neq j$ and $X_i^n \perp Y^n$ where $\perp$ means "statistical independence" $i, j = 1, \dots m - 1$ and $n \geq 0$, which implies:

- $E[X_rY] = E[X_r]E[Y]$.
- $E[X_rY^2] = E[X_r]E[Y^2] = E[X_r][Var(Y) + E[Y]^2]$
- $cov(Y, X_rY) = E[X_rY^2] - E[X_rY]E[Y] = E[X_r]Var(Y)$
- $cov(X_iY, X_jY) = E[X_iX_jY^2] - E[X_iY]E[X_jY] = E[X_i]E[X_j]Var(Y)$ if $i \neq j$
- $cov(X_iY, X_iX_jY) = Var(X_i)E[X_j]Var(Y) + Var(X_i)E[X_j]E(Y)^2 + E[X_i]^2E[X_j]Var(Y)$ if $i \neq j$
- $cov(Y, X_iX_jY) = E[X_i]E[X_j]Var(Y)$

For $m = 1$, we have:

$$Var(Z_m) = Var(Y)$$

For $m = 2$:

$$Var(Z_m) = Var[(1 - X_1)Y] = Var(Y - X_1Y) = Var(Y) + Var(X_1Y) - 2cov(Y, X_1Y)$$
$$= Var(Y) + Var(X_1Y) - 2E[X_1]Var(Y)$$

For $m = 3$:

$$Var(Z_m) = Var[(1 - X_1)(1 - X_2)Y] = Var(Y - X_1Y - X_2Y + X_1X_2Y) = Var(Y) + Var(X_1Y) + Var(X_2Y) + Var(X_1X_2Y) - 2cov(Y, X_1Y) - 2cov(Y, X_2Y) + 2cov(Y, X_1X_2Y) + 2cov(X_1Y, X_2Y) - 2cov(X_1Y, X_1X_2Y) - 2cov(X_2Y, X_1X_2Y) = Var(Y) + Var(X_1Y) + Var(X_2Y) + Var(X_1X_2Y) - 2E[X_1]Var(Y) - 2E[X_2]Var(Y) + 2E[X_1]E[X_2]Var(Y) + E[X_1]E[X_2]Var(Y) - 2\{Var(X_1)E[X_2]Var(Y) + Var(X_1)E[X_2]E(Y)^2 + E[X_1]^2E[X_2]Var(Y)\}10 - 2\{Var(X_2)E[X_1]Var(Y) + Var(X_2)E[X_1]E(Y)^2 + E[X_2]^2E[X_1]Var(Y)\}$$

Therefore, the general formula for $Var(Z_m)$ is complicated to derive and we used simulation-based methods to estimate the variance of Z_m in our study