Infection Aware Nurse Staffing Using Random Graphs with Hidden Health Status

Buyun Li • Joanthan E. Helm • Kurt M. Bretthauer
Operations and Decision Technologies, Kelley School of Business, Indiana University,
libu@iu.edu • helmj@iu.edu • kbrettha@iu.edu

During outbreaks of infectious diseases, hospitals encounter significant challenges in making well-informed nurse staffing decisions. The dilemma during these outbreaks involves a simultaneous increase in inpatient admissions during disease outbreaks and a notably elevated rate of nurse absenteeism caused by infections. The unobservable nature of nurse infection time, incubation period, and number of nurses infected but yet to show symptoms adds complexity to understanding when and how nurses are infected. Lack of this critical information restricts hospital managers from implementing effective and informed operational strategies and staffing plans, limiting their ability to proactively address the staffing crisis during an outbreak. We develop a dynamic random graph model with hidden nurse health status to examine the interplay between staffing policies and infection transmission dynamics. Our model extends existing random graph frameworks by incorporating nurse health status (healthy, incubation, symptomatic) as a latent variable that is endogenously linked to the evolution of disease transmission networks. Within this framework, we design an estimation procedure that maps nurse characteristics to disease transmission rates across patient-to-nurse, nurse-to-nurse, and community-to-nurse interactions. This approach enables dynamic tracking of infection sources, locations, and timing. Using data from the IU-Health hospital system during the COVID-19 pandemic, we perform counterfactual analyses to assess the effectiveness of mitigation and staffing policies aimed at protecting nurses from infections. We find that hospitals can reduce nurse absenteeism due to infection by up to 25% through improved staffing levels and workload management. Furthermore, when establishing dedicated units for the care of infectious patients, simply isolating infected patients is insufficient; it is crucial to assign a fixed group of nurses exclusively to these patients to minimize cross-infection.

Key words: In-Hospital Infection; Nurse Staffing; Healthcare Resilience; Random Graph Model

1. Introduction

Over the last 25 years, healthcare workers have battled a wave of severe outbreaks of infectious diseases. Amid these outbreaks, healthcare workers have faced the difficult dual challenge of providing critical care to infected patients while protecting themselves against the risk of infection. This dual challenge has, in turn, caused serious staff shortages in hospitals, as the influx of infected patients increases dramatically during these outbreaks. At the same time, healthcare workers experience higher infection rates than the general public. During the 2003 SARS outbreak, healthcare workers were found to be 18 times more susceptible to infection compared to the general population, accounting for almost one third of all infections (Peck et al. 2004). In the 2009 swine flu pandemic, healthcare workers were 11 times more likely to be infected (Lietz et al. 2016). Similarly, in the 2012 Middle East respiratory syndrome coronavirus outbreak, healthcare workers represented approximately 20% of infections, with a 14 times higher likelihood of being infected than the general public (Suwantarat and Apisarnthanarak 2015, Elkholy et al. 2019). The 2013-2016 epidemic of Ebola virus disease in West Africa, as well as the 2015 epidemic of Zika virus disease, both resulted in substantial in-hospital infections and mortality among healthcare workers, which consequently exacerbating shortages within the healthcare workforce (Suwantarat and Apisarnthanarak 2015). Most recently, during the COVID-19 pandemic, healthcare workers were 12 times more likely to be infected by COVID-19, leading to 93% of hospitals experiencing understaffing during the pandemic, a nearly 50% jump from the pre-pandemic period (Xiao et al. 2020, Nguyen et al. 2020, Kayser 2023). In addition to these significant outbreak events, healthcare workers have consistently been more susceptible to contracting seasonal infectious diseases such as influenza and stomach flu, which infect healthcare workers five times more frequently than the general public (Pereira et al. 2017). According to Pereira et al. (2017), on average, around 15% of healthcare workers are absent from work due to influenza and stomach flu infection during their active seasons, which causes a consistent issue for healthcare workers staffing. The elevated infection rates, coupled with the consequent overwhelming workload, have caused burnout in approximately 50% of healthcare workers, contributing to a significant exodus from the profession in the aftermath of outbreaks (Sullivan et al. 2022). Our focus is specifically on Registered Nurses and associated infections, as nurses represent the majority of healthcare workers and the group most frequently impacted by infection-induced absenteeism.

Although the impact of clinical factors such as disease contagiousness and hospital patient census on hospital infections among nurses is widely recognized, operational factors and the experience of nurses taking

care of infectious patients also emerge as crucial contributors. The operational factors, such as workload levels set in the hospital, scheduling routines, and the mix of infectious and non-infectious patients, also play an important role in nurses' exposure to the disease. For example, according to a meta-study of the COVID-19 pandemic, 73% of nurses think that they were infected during the chaos caused by the heavy workload in the hospital, and more than 90% of nurses believe that working back-to-back shifts increased their chance of infection (Firouzkouhi et al. 2022). Interestingly, research has also indicated that the experience of nurses protecting themselves while caring for infectious patients is significant in the context of hospital infections. According to a review by Jackson et al. (2023), it is observed that the experience over time of nurses caring for patients with COVID-19 can eventually have a dual effect, helping reduce the likelihood of contracting the virus not only in the hospital setting but also in the community. This effect comes from the learning effect on the use of appropriate personal protective equipment (PPE), the requirement of sterilization procedures, and the avoidance of high-risk areas and activities.

In response to disease outbreaks, hospitals have formulated a wide range of mitigation and staffing policies. However, understanding the effectiveness of these policies individually or collectively poses a challenge. During major pandemics such as Swine Flu and COVID-19, hospitals established dedicated departments to handle cases (Anand et al. 2012, Patel et al. 2009, Nguyen et al. 2020). For highly fatal diseases such as SARS and Ebola, hospitals instituted strict protocols for personal protective equipment (PPE) when dealing with affected individuals. In addition, hospitals developed new sick leave policies for nurses and introduced vaccination compliance policies as part of efforts to contain the spread of diseases. However, it is important to acknowledge that the development and deployment of mitigation strategies have largely taken place ad hoc, with significant variations between countries, states, counties, and hospitals, and therefore potentially lacking generalizability. Consequently, there is a clear need for a rigorous analysis that can comprehensively assess the cumulative impact of the diverse spectrum of interventions that are being enacted. Our paper provides a systematic approach to comprehensively model the interaction between disease transmission and staffing decisions, which enables a more coherent and evidence-based strategy evaluation and facilitates a swift and comprehensive response to challenging operational decisions.

Another crucial factor that adds complexity to the assessment of staffing and mitigation policies is hospitals' limited knowledge of the transmission pathways of infections, which often eludes direct observation. The distribution of these sources responsible for in-hospital infections among nurses, effectively dictating

where and when nurses contract diseases, continues to be a subject of ongoing debate. Taking COVID-19 as an example, based on contact trace data, some infectious disease studies have found that a significant proportion of nurses, approximately 40%, contracted the disease from other infected colleagues, compared to approximately 38% from patients and 22% from community and nonhospital settings, as reported by Nguyen et al. (2020) and Al Maskari et al. (2021). In contrast, the Mayo Clinic and another subsequent study argued that the majority of nurses contracted the virus within the community (Wong 2020, Harith et al. 2022). The disparity in these studies with respect to the source of infection can be attributed to the heterogeneous and unobservable nature of the disease's infection time and incubation period. Because it is impractical to determine the specifics of when a healthcare worker is infected and the length of the incubation period during which a nurse can infect other colleagues, the reliability of determining the source of infection solely on the basis of contact tracing is questionable (Kwok et al. 2019). Furthermore, there is a lag in information between the hospitals and the nurses, i.e., hospitals are only informed of nurses exhibiting symptoms when a nurse is absent from work or a very short time before when nurses are scheduled to work. To accurately estimate the source of infection, we develop a probabilistic approach that changes the variable of the study to be the time range of a nurse showing symptoms instead of the exact date of the nurse's infection. In doing so, we can take into account the random and unobservable infection time, incubation time, and information time lag.

Our study is among the first to contribute to the resilience of healthcare systems by addressing the critical issue of infection-aware nurse staffing amid disease outbreaks. We introduce a dynamic random graph model that integrates nurses' hidden health statuses, thereby capturing the intricate interplay between disease transmission dynamics within hospitals and staffing decisions. This model provides detailed and tractable insights into the pathways of infection—from interactions with infectious patients, transmission among nursing staff, and exposures originating in the community—facilitating a systematic evaluation of diverse staffing strategies and mitigation policies aimed at reducing nurse infections. Furthermore, the granular perspective on transmission dynamics offered by our counterfactual analysis clarifies the mechanisms underlying infection-aware staffing and mitigation strategies, providing broadly applicable insights for healthcare operations during outbreaks.

Building on the existing literature on dynamic networks with hidden variables, our paper extends this work by incorporating the endogeneity between these variables and the network structure. Specifically, we

examine scenarios where hidden variables—such as health statuses in disease transmission networks—are endogenously determined by the network structure and, in turn, influence its future evolution. Existing studies on dynamic networks with hidden variables often assume that these variables are exogenous and independent or fully observable. However, this endogenous interaction is critical for accurately modeling disease spread, but it remains underexplored in prior work. One of our key contributions is to define this endogenous relationship within a dynamic network framework and to develop an estimation procedure that addresses the resulting complex dependencies. This approach enables consistent estimation of model parameters, providing a more realistic and broadly applicable method for analyzing disease transmission networks and other network settings where hidden variables and network structures are interdependent.

In addition, our counterfactual analysis extends the broader nurse staffing literature by explicitly linking operational decisions to infection dynamics in ways that traditional models of absenteeism have seldom addressed. Specifically, we demonstrate that adjusting workload levels, implementing on-call capacity, and designating specialized units for infectious patients can significantly affect both the overall incidence of nurse infections and the pathways through which those infections spread. Such findings highlight how disease outbreaks, rather than being treated as exogenous shocks, can be systematically modeled as an integral part of staffing decisions. This perspective complements studies that focus on endogenous absenteeism due to factors such as workload or burnout by showing that nurse infections themselves can be shaped—or mitigated—through targeted operational policies. By unveiling how staffing configurations drive infection risk and consequent workforce disruptions, our framework provides a novel analytical lens for designing more resilient nurse staffing systems during disease outbreaks.

The classic random graph model (Erdős and Rényi 1959) is a widely recognized approach to modeling disease transmission (see Section 2 for a detailed review). In this framework, individuals are represented as nodes and disease transmission is modeled with stochastic edges between nodes, with probabilities determined by node-specific features. Importantly, the existing literature modeling disease transmission as a random graph assumes that the health status, i.e., being healthy, incubation or symptomatic, is known to the decision maker at all times. In our context, similar to the classic approach, we model patients, nurses, and community infection sources as nodes, and disease transmissions to nurses as edges. However, we relax the assumption that the health status of the nurse is observable at all times. We assume that the health status of nurses is only revealed when they are absent from work, at which point we know that they are symptomatic.

The hidden nurse health status provides a realistic representation and estimate of the disease transmission process. Instead of making assumptions on the specific point in time when a nurse is infected or symptomatic, and how many nurses are in the incubation period infecting other nurses, we use a probabilistic approach based on intervals of time to systematically capture the nurses' hidden health status and the transitions of their health status, e.g., from healthy to infected, infected to symptomatic. With hidden nurse health status, our model captures the effects of random incubation periods, unobservable infection times, infection sources, and time lags between the onset of symptoms and actual absenteeism. Using probabilistic inference and the Expectation-Maximization (EM) method, we develop an estimation procedure for this dynamic random graph model with hidden status. This approach gives us tractability for estimating disease spread and nurse status at a granular level, facilitating a more comprehensive evaluation of infection mitigation policies and the corresponding staffing strategies.

We partner with IU-Health, one of the nation's largest healthcare systems, to implement and test our framework with extensive real-world data, spanning 18 hospitals in various sociodemographic regions. This collaboration not only provides a high-fidelity environment for parameterizing and validating our infection-aware staffing model, but also ensures that our findings reflect on-the-ground operational realities. In our analysis of one major facility, we find that approximately 70% of nurse infections originate within the hospital (split almost evenly between patient-to-nurse and nurse-to-nurse transmissions), while the remaining 30% stem from the community. Our counterfactual simulations show that refining workloads and scheduling can reduce nurse infections by up to 25%, demonstrating substantial potential for mitigating staff shortages. Moreover, introducing on-call nurses—who can rapidly cover unexpected absences—lower infection-driven absenteeism by an additional 8%. We also see that dedicating a fixed nursing team to an infectious cohort is significantly more effective than merely isolating infected patients, and higher vaccination uptake offers stronger staff protection than more effective yet lower-compliance vaccination regimens. These insights, validated through close collaboration with our healthcare partners, underscore the vital role of operational policy levers in reducing infections and preserving frontline workforce capacity.

The remainder of this paper is organized as follows. In section 2, we review several relevant streams of research and discuss our contributions to this literature. Section 3 introduces the dataset and reports a model-free comparison of key features between nurses with and without absenteeism caused by infection. Section 4 introduces our stochastic network model of inpatient hospital units, incorporating a random graph

component to model disease transmissions, and we specify our chosen estimation methodologies. In Section 5, leveraging the estimated parameters, we analyze hospital operational strategies and report the most effective policies.

2. Literature Review

In the previous section, we introduced literature related to pandemics and their impact on nurse absenteeism. Here, more generally, we focus on how a variety of endogenous variables may impact nurse absenteeism. In addition, we review disease transmission models, random graph network models, and delay time models related to inspection and maintenance. Nurse staffing in the presence of endogenous absenteeism provides the closest context for our model of nurse absenteeism caused by infections. Green et al. (2013) study nurse staffing decisions taking into account the endogenous issue of how workload impacts nurse absenteeism. Their work extends early literature focused on staffing decision in service environments that either ignores or treats absenteeism as an exogenous factor (Easton and Goodale 2005, Whitt 2006, He et al. 2012). Wang and Gupta (2014) explore a similar problem, where additional factors such as unit type and culture are used to predict nurse absenteeism. Additionally, Kluger et al. (2020) investigates the impact of extending short nurse shifts in addressing nurse absenteeism caused by COVID-19. However, Kluger et al. (2020) predict nurse absenteeism solely based on exogenous parameters, overlooking the potential for nurse infection from other colleagues or the community. Our study contributes to the literature by introducing a model that accounts for nurse infection from all sources, including patients, other nurses, the community, and their interactions within a dynamic random graph network model. This framework enables us to capture the system dynamics between nurse staffing decisions and the embedded disease transmission network.

The body of literature focusing on mathematical modeling of infectious disease has been growing rapidly; see Vynnycky and White (2010) for a thorough introduction. These models can be categorized into three streams: statistical forecasting models, i.e. r_0 models, curve-fitting or time-series methods (Shakeel et al. (2021)), compartmental models (Beckley et al. (2013)), i.e. SIR models, and random graph network models (Keeling and Eames (2005)). Statistical forecasting models and SIR models are limited in their ability to model nurse-individual and operational characteristics' impact on disease spread, and these models often build on population uniform mixing assumptions. Therefore, we choose the random graph model for its ability to capture detailed nurse characteristics and operational features that are related to the infection

and its capability to differentiate infections between different sources with dynamic rates based on these features. In this stream, Eames and Keeling (2002) model sexually transmitted diseases (STDs) based on a social network; Blanchard et al. (2005) model the spread of STDs on a bipartite graph; Drakopoulos and Zheng (2017) study network effects and their implication for infectious disease control. Venkatraman et al. (2021) provide a detailed survey on the applications of random graph networks in modeling infectious disease transmission. However, virtually all applications of the random graph network model assume that the transmissions among nodes are observable and the health status of the nodes is known at all times for all stages. We contribute to the infectious disease literature by relaxing this fundamental assumption. This is a critical advancement because the unobservable disease-spreading events and the hidden infectious status of patients during incubation periods are the fundamental factors contributing to the complexity and dynamics of infectious disease modeling.

In addition to the contributions to the disease modeling literature just discussed, our random graph network model with hidden status extends the literature on network models. Erdős and Rényi (1959) is often cited as the original reference for random graph network models. In Erdős and Rényi (1959), a basic random graph is defined as a graph with n nodes that is constructed by connecting each pair of nodes with an edge independently with a probability p, which is known as the Erdős–Rényi (ER) model. Following this definition, contemporary literature assigns characteristics to each node and models the connecting probability p as functions of those characteristics; see (Burt 1980, Carrington et al. 2005, Lewis 2011) for a thorough introduction. Random graph network models, in recent years, also have important applications in the operations literature, e.g. supply chain management (Göttlich et al. 2005), social networks (Scott 2012), and healthcare management Brunson and Laubenbacher (2018). More recently, dynamic random graphs with latent node characteristics, in which node characteristics follow a stochastic process and evolve through time, have gained increasing traction. Hoff et al. (2002) first introduced a static latent space model (LSM) where the position of each node follows a distribution. The position of node pairs affects the probability of the presence of an edge connecting these nodes. Sarkar and Moore (2005) extend the static LSM to a dynamic LSM by incorporating a stochastic process to model the evolution of positions of nodes. In parallel to LSM, Nowicki and Snijders (2001) first studied the effect of latent clustering of the nodes in a network by offering a static stochastic block model (SBM). In Nowicki and Snijders (2001), nodes are partitioned into multiple blocks according to an exogenous probability distribution. The clustering of the nodes (known as membership)

then affects the edge presence among node pairs. Yang et al. (2011) extend the static SBM to a dynamic SBM by modeling membership as stochastic processes rather than probability distributions. More recently, Friel et al. (2016) offer an LSM with bipartite clustering of the nodes that can be considered as a bridging model linking LSM to SBM. Kim et al. (2018) survey the evolution and extension of the above models in great detail. Literature in this stream builds on the assumption that the latent characteristics follow a known distribution or transition probability that is independent from the edge presence among nodes or any other node characteristics. However, this assumption does not apply to disease transmission settings where edge presence denotes a transmission event and latent characteristics represent the health status of a nurse. From one direction, health status affects the dynamics of the disease transmission, i.e. edge presence is only possible if it is from an infectious node to a healthy node; from the other direction, health status is also determined by the edge presence, i.e. if there is a transmission event from an infectious source to a nurse, that nurse cannot remain healthy. While the endogenous relationship between edge presences and node latent status significantly complicates the dynamics and estimation of the random graph network model, it is a necessary feature to study disease transmission in this capacity. To the best of our knowledge, we are the first to consider the set-up of a random graph network in which the latent status of the nodes and the edge presence relationship are endogenously determined. Such a setting is also useful in other contexts; e.g. in LSM modeling of social networks, individuals are more likely to move to a location where their connections reside than a location where they have no connections.

Estimation of nurse infections with latent incubation time can also be conceptualized as a delay time model within the framework of machine maintenance and inspection settings, as extensively surveyed by Wang et al. (2011), Wang (2012). In this analogy, we equate a nurse infection, an inherently unobservable event, to a latent machine defect leading to a breakdown in the future. Upon the end of the incubation period, an infected nurse shows symptoms, akin to a machine breakdown event that remains unobservable to the decision-maker. Each nurse shift parallels a machine inspection, where absenteeism resulting from infection is equivalent to the event of discovering the machine in a failure state. Recently, generalized delay time modeling has gained increasing attention in the operations literature. For example, in healthcare Chan et al. (2012) and Liu et al. (2018) use delay time models to investigate patient readmission time in relationship to discharge and post-discharge checkup decisions; in an industrial setting, Zhao et al. (2015) demonstrate the practical application of delay-time models; in an environmental setting, Santos et al. (2021) apply delay time models

to re-use of items. Our paper is the first to use a model that could be conceptualized as delay-time modeling to provide probabilistic analysis in the context of nurse infection and absenteeism. Our problem necessitates several extensions to the classical delay time model (Baker and Wang 1993): (1) State-dependent delay time: unlike the classical model, we recognize that the delay in absenteeism is contingent upon characteristics of nurses, units, and hospitals. (2) Machine defects are correlated and have cascading effects upon each other: infection of a nurse may lead to more infections of other nurses within the same unit due to nurse-to-nurse disease transmission. (3) Inspection cannot reveal the defect in machines and the inspection time, which corresponds to nurse shift schedules, is heterogeneous among nurses and unevenly spaced. Our model directly contributes to classical delay time models by incorporating the above extensions. We also offer an estimation procedure that handles these unique extensions.

3. Data Overview

Our analysis relies on two datasets: (1) nurse staffing and absence data and (2) hourly patient-flow census data. Both datasets were obtained from IU Health and cover the period from March 11, 2020, when the first COVID-19 inpatient was observed at IU Health, to July 21, 2020, encompassing the transition from the 'first wave' to the 'second wave' of the pandemic. The nurse staffing data serves as the basis for our outcome variable, which indicates whether a nurse missed a shift due to sick leave resulting from contracting COVID-19 (referred to as 'nurse absenteeism' throughout the paper). Meanwhile, the patient-flow data is utilized to calculate various covariates used in our predictions of nurse absenteeism due to COVID-19, including, for example, the total number of COVID-19 patients admitted each day and the patient-to-nurse ratio. Below, we provide a detailed description of both datasets and outline our sample structure, as well as our model-free comparison of statistics between nurses who were absent from work due to COVID-19 and nurses who were not absent.

3.1. Datasets Introduction

The staffing dataset encompasses individual-level records for more than 5,000 nurses across 18 hospitals within our collaborator's healthcare network. This comprehensive dataset contains details for each healthcare worker, including their role (e.g., registered nurse, lead nurse), whether they are involved in direct patient care or have indirect responsibilities (e.g., nurse unit managers, assumed to have indirect care), the scheduled shift date, and the corresponding hospital unit ID. Additionally, for every scheduled shift, the dataset

	Nurses with Absenteeism			Nurses without Absenteeism		
Metrics	Med-Surg	PCU	ICU	Med-Surg	PCU	ICU
Patient-Nurse Ratio	4.23 (+7%)	3.02 (+9%)	1.51 (-0.2%)	3.95	2.78	1.54
COVID-19 Interaction	37.31 (+25%)	30.11 (+4%)	32.88 (+12.5%)	29.51	28.82	29.20
Nurse Interaction	5.02 (+14.4%)	5.66 (+0.2%)	11.12 (+20%)	4.37	5.575	9.28
Cumulative COVID Interaction	337.07 (-61%)	444.05 (-41%)	433.25 (-41%)	874.03	822.85	748.25

Table 1 Comparative descriptive statistics of nurses with vs. without absenteeism due to COVID-19 infection

provides information on whether each nurse showed up for their shift. If the nurse attends the shift, the data records the shift start and end times. Otherwise, if the nurse is absent from the shift due to COVID-19 infection, the dataset uses a distinct flag to indicate missed shifts attributed to nurses who contracted COVID-19 and subsequently required sick leave. This specific code serves as the basis for our primary outcome variable, which measures nurse absenteeism resulting from COVID-19 infection. Our study focuses on nurses directly interacting with patients in adult inpatient units, excluding nurses working in laboratory settings, administrative roles, or neonatal intensive care units (NICU). Consequently, we have detailed scheduling data for 3,622 nurses who worked a total of 123,837 shifts. Among these 3,622 nurses, 468 (13%) reported sick leave due to COVID-19 infection in the 4-month period.

The patient census dataset comprises approximately 30,000 patient visits, each documented at an hourly level. For every hour of every patient visit, the dataset includes the following information: patient location, recorded as the unit ID, the level of care provided (distinguishing between critical care and non-critical medical/surgical care), whether the patient tested positive for COVID-19, and whether the patient required a ventilator. Throughout this period of time, IU Health admitted a total of 1,605 COVID-19 patients. The patient census data serves as the basis for calculating covariates associated with patient workload within a hospital unit on any given day. These covariates are linked to the respective nurses' shifts by matching the shift time and unit ID. In addition, we establish connections among nurses themselves by matching shift times and units worked.

3.2. Nurses absenteeism caused by infection

Table 3 in the Appendix summarizes and defines all the features we extracted from our dataset by combining the patient census data and the nurse working data. Table 1 presents comparisons of key features between nurses affected by absenteeism related to COVID-19 and those not affected, highlighting differences in magnitude and percentage (in parenthesis). Among nurses who were infected, we observed that they encountered a higher patient-to-nurse ratio, cared for more infectious patients per shift, and had more interactions with

fellow nurses during their shifts. Based on a discussion with our hospital collaborator, another important factor that affects nurse infection is the experience of a nurse caring for infectious patients. Nurses with less experience caring for infectious patients are more vulnerable to infections as they may lack the necessary skills and knowledge to adequately protect themselves from infected patients or colleagues, adhere to safety protocols, and effectively manage the complexities of treating infected patients while protecting themselves. We use the cumulative interaction time of a nurse caring for COVID-19 patients to measure this experience. It is important to note that the experience difference between the two groups of nurses is inherently endogenous to the infection: Infected nurses work fewer hours throughout the data since they are removed from the workforce after infection. Such an endogenous relationship is known as right-censoring. To correct for right-censoring and rigorously identify a relevant and robust feature set to predict nurse infections, we employ a feature selection procedure using a Cox model to select the important variables for predicting infections from the data we extracted. This procedure is detailed in Section 4.3.

4. Hierarchical Random Graph Network Model with Endogenous Hidden Health Status.

We develop a hierarchical random graph network model with latent health status to characterize the dynamics of disease transmission, nurse staffing decisions, and their interdependencies, as detailed in Section 4.1. In this framework, nurses, patients, and community infection sources are represented as nodes, while transmission events are modeled as edges. A key feature of our model is that the health status of nurses, specifically whether they are healthy, infectious (during the incubation period), or symptomatic, remains unobservable until a nurse reports infection by calling in sick for an upcoming shift. This latent status introduces a fundamental distinction between our approach and existing network models on disease transmission. The consideration of hidden health status is essential to accurately and correctly model disease transmission in a network setting. Because disease transmission is unobservable and incubation periods vary between individuals, both the health status of a nurse and the transitions of health status, such as from healthy to incubation, remain unobservable. As a result, the direction of spread of the disease among nurses is also unobservable: infectious nurses serve as sources of transmission, while healthy nurses remain susceptible to infection. The unobservability of health status and transmission direction fundamentally impacts the structure and dynamics of the disease transmission network, as it remains unknown which nurse is spreading the disease and which nurse is susceptible at any given time. In addition, in real-world healthcare operations,

not only is the timing of a nurse's infection unobservable, but the onset of symptoms also remains unknown. Rather than identifying the exact moment that a nurse becomes symptomatic, hospitals can only learn this status when the nurse calls in sick for an upcoming shift. These factors further complicate the dynamics of disease transmission and the estimation of the underlying transmission network.

To address these complexities, we introduce a novel estimation procedure that explicitly incorporates the latent nature of nurse infections, as elaborated in Section 4.2. In contrast to conventional delay-time models that focus on discrete time points, our methodology emphasizes intervals over which infection events occur. Estimation of the disease transmission network under hidden health status requires reframing the inferential objective: rather than pinpointing the exact day that a nurse contracts the disease, we determine whether the onset of symptoms occurs within the interval spanning the nurse's last recorded healthy shift and their subsequent notification of sick leave. This shift in perspective allows for a more robust characterization of disease transmission pathways in the presence of latent health states.

Our estimation procedure performs well when applied to real data from IUH during the first wave of COVID-19, see Section 4.4. With the adoption of our hierarchical random graph model, we can provide insight into the composition of transmissions from different sources (nurse, patients, and the community outside the hospital) during that period for a large hospital.

4.1. Model Overview: Hierarchical Random Graph Network Model with Endogenous Hidden Status

We consider a dynamic random graph, $G = \{G_1, \ldots, G_T\}$, where each G_t for $t = 1, \ldots, T$ represents a snapshot of the graph over time period t. Each graph G_t is defined as $G_t = (\mathcal{K}, \chi_t)$, consisting of a set of nodes $\mathcal{K}$ and an edge adjacency matrix $\chi_t = (x_{ijt})_{(i,j)\in\mathcal{K}}$. Here, x_{ijt} is a binary variable that takes the value 1 if an edge exists from node i to node j in period t and 0 otherwise. In the random graph, we model nurses, patients, and community transmission sources as nodes and disease transmissions as edges between nodes. We provide details on the nodes and edges specification:

The set of nodes is defined as $\mathcal{K} := \mathcal{N} \cup \mathcal{P} \cup C$, where $\mathcal{N}$ represents the set of nurse nodes, labeled as $1, \ldots, N$; $\mathcal{P}$ represents the set of patient nodes, labeled as $1, \ldots, P$; and C denotes the community infection source node. Consequently, the graph G_t for $t = 1, 2, \ldots T$captures the interactions between all $K = N + P + 1$ nodes over time.

Each nurse node $j \in \mathcal{N}$ has two main attributes: a *transmission feature* (d_{jt}) and a *health status* (h_{jt}). The transmission feature captures factors influencing disease spread, such as nurse j's average patient-to-nurse

ratio in period t and the nurse's level of experience. Meanwhile, h_{jt} tracks the nurse's health state over time. For patient nodes and the community infection source node, i.e., $j \in \mathcal{P} \cup C$, only the health status (h_{jt}) is monitored.

We denote the health statuses of all nodes at time t by the vector $H_t = (h_{jt})_{j\in\mathcal{K}}$. A nurse's health status can be one of three states: *healthy*, *infectious*, or *symptomatic*. At the start of the time horizon, all nurses are assumed to be healthy, i.e. $(h_{j1})_{j\in\mathcal{N}} = \text{healthy}$. However, a nurse's status remains hidden until the nurse calls in sick, which only occurs if a symptomatic nurse is scheduled to work. By contrast, patient health statuses are observable, since admitted patients are tested and classified as either *infectious* or *non-infectious*. The community infection source node C is considered *infectious* throughout the horizon, modeling potential disease transmission to nurses from outside the hospital.

Within the hospital, transmissions are assumed to occur solely within designated units. For each period t, the unit assignments $U_{1t}, \ldots, U_{Wt} \subseteq \mathcal{K}$ define subsets of the node set $\mathcal{K}$, where W is the total number of units. These assignments capture which nurses and patients occupy each unit over every period. To incorporate disease transmission outside the hospital, we introduce a "community unit," $U_C = \mathcal{N} \cup C$, which includes all nurses plus the external infection source C. Given these unit assignments, we track the interaction time among nodes within the same unit. For $i \in \mathcal{K}, j \in \mathcal{N}$, and $t \in \{1, \ldots, T\}$, let s_{ijt} represent the duration that node i and nurse j spend together in the same unit during period t, that is, $\{i, j\} \in U_{kt}$ for some $k \in \{1, \ldots, W\}$.

For each period $t \in \{1, \ldots, T\}$, edges in the network are represented by a $K \times K$ adjacency matrix $\chi_t = (x_{ijt})_{(i,j)\in\mathcal{K}}$. An entry $x_{ijt} = 1$ denotes the occurrence of a disease transmission event from node i to node j during period t, while $x_{ijt} = 0$ indicates no transmission. If $x_{ijt} = 1$, nurse j transitions from *healthy* to *infectious* in the next period (i.e., $h_{j,t+1} = \textit{infectious}$). After a random incubation phase in the *infectious* state, nurse j then becomes *symptomatic*.

An edge between two nodes requires that they are *connected*. Specifically, node i is connected to node j during period t, denoted $i \overset{t}{\sim} j$, if:

1. Node i is infectious (whether a nurse, a patient, or the community node), i.e. $h_{i,t} = \textit{infectious}$.
2. Node j is a healthy nurse, i.e. $h_{j,t} = \textit{healthy}$ for $j \in \mathcal{N}$.
3. Nodes i and j occupy the same unit during period t, or i is the community node C. Formally,

$$\big(\exists\, k \in \{1, \ldots, W\} : i, j \in U_{kt}\big) \quad \text{or} \quad (i = C).$$

The notation $i \overset{t}{\sim} j$ thus indicates a *possible* transmission path from an infectious node i to a healthy nurse j. Conversely, $i \overset{t}{\nsim} j$ means such a transmission path is not possible. In short, *connectivity* defines the potential for an edge to form, whereas $x_{ijt} = 1$, the presence of an edge between i and j denotes an *actual* disease transmission event.

We assume that the transmission time from an infectious node i to a connected healthy node j is independent of all the transmission times between any other infectious nodes and the healthy node j. To account for different sources of transmissions, we further divide the feature vector for each nurse, d_{jt} for $j \in \mathcal{N}$, into three subsets: nurse-to-nurse, patient-to-nurse, and community-to-nurse transmission features, denoted by $d^n_{i,t}, d^p_{i,t}, d^c_{i,t}$, respectively. We model the conditional probability of edge presence with a hierarchical exponential-linear model with transmission-type-specific parameters sets $\theta = (\theta^k_1, \theta^k_2)_{k\in\{n,p,c\}}$:

$$P(x_{ijt} = 1 | H_t, d_{it}, s_{ijt}, i \overset{t}{\sim} j) = 1 - e^{\gamma_{ijt} s_{ijt}} \tag{1}$$

$$\gamma_{ijt} = \begin{cases} \theta^n_1 + \theta^{n\intercal}_2 d^n_{j,t} & \text{if } i \in \mathcal{N} \\ \theta^p_1 + \theta^{p\intercal}_2 d^p_{j,t} & \text{if } i \in \mathcal{P} \\ \theta^c_1 + \theta^{c\intercal}_2 d^c_{j,t} & \text{if } i = C \end{cases}, \tag{2}$$

where γ_{ijt} is the disease transmission rate from an infectious node i to a node j during day t. We choose a hierarchical exponential-linear model because it provides analytical tractability. To see this, let $X_{jt} = 1$ denote the event of nurse j being infected in period t and $X_{jt} = 0$ otherwise. Given the exponential form, for a healthy nurse j the conditional probability $P(X_{jt} = 1 | H_t, d_{j,t}, s_{ijt} : i \overset{t}{\sim} j)$ is given by:

$$\begin{aligned} P(X_{jt} = 1 | d_{j,t}, s_{ijt} : i \overset{t}{\sim} j) &= 1 - \prod_{i\in\mathcal{K}} P(x_{ijt} = 0 | h_{jt} = \textit{healthy}, H_t, d_{j,t}, s_{ijt} : i \overset{t}{\sim} j) \\ &= 1 - e^{\sum\limits_{i\in\mathcal{K}} \gamma_{ijt} s_{ijt}} \end{aligned} . \tag{3}$$

Note that γ_{ijt}, only depends on the type of node i, therefore, instead of tracking all individual pairs of transmission rates, it is sufficient to track transmission rates from the three sources to node j by a vector: $\gamma_{jt} = \begin{bmatrix} \gamma^n_{jt} \\ \gamma^p_{jt} \\ \gamma^c_{jt} \end{bmatrix}$, which stand for transmission rates from nurse-to-nurse, patient-to-nurse, and community-to-nurse sources. Based on equation (2), we specify the mapping from feature set $d_{j,t}$ to γ_{jt} as :

$$\gamma_{jt} = \begin{bmatrix} \gamma^n_{jt} \\ \gamma^p_{jt} \\ \gamma^c_{jt} \end{bmatrix} = \begin{bmatrix} \theta^n_1 + \theta^{n\intercal}_2 d^n_{j,t} \\ \theta^p_1 + \theta^{p\intercal}_2 d^p_{j,t} \\ \theta^c_1 + \theta^{c\intercal}_2 d^c_{j,t} \end{bmatrix}. \tag{4}$$

Let $S_{jt} = \begin{bmatrix} S^N_{jt} \\ S^P_{jt} \\ t'_j \end{bmatrix}$ denote the interaction time vector for nurse j during day t, where $S^N_{jt} = \sum\limits_{i\in\mathcal{N}: i \overset{t}{\sim} j} s_{ijt}$ is the total connection time between nurse j and infectious nurses during day t, $S^P_{jt} = \sum\limits_{i\in\mathcal{P}: i \overset{t}{\sim} j} s_{ijt}$ is the total connection

time between nurse j and infectious patients during day t, and t'_j is the total time nurse j spent in community during day t. With the above notation, equation (3) simplifies to

$$P(X_{jt}=1|H_t, d_{j,t}, s_{ijt} : i \overset{t}{\sim} j) = 1 - e^{\gamma_{jt}^{\intercal} S_{jt}}. \tag{5}$$

For ease of exposition, throughout the rest of our analysis, we refer to equation (4), the feature set to transmission rate mapping, as $\phi(\cdot|\theta) : d_{jt} \xrightarrow{\theta} \gamma_{jt}$ and equation (5), the transmission rate to infection probability mapping, as $\psi(\cdot|S_{jt}) : \gamma_{jt} \xrightarrow{s_{jt}} 1 - e^{\gamma_{jt}^{\intercal} S_{jt}}$. Finally, we denote the composite mapping from the feature set to infection probabilities as:

$$\phi \circ \psi(\cdot|S_{jt}, d_{jt}) : d_{jt} \xrightarrow{\theta, S_{jt}} 1 - e^{\gamma_{jt}^{\intercal} S_{jt}}. \tag{6}$$

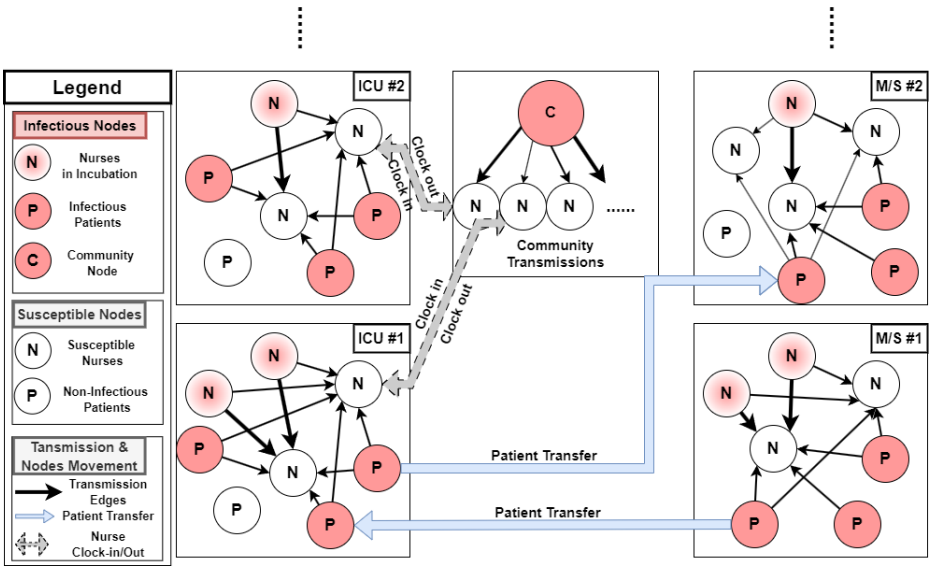


Figure 1 **Diagram for the random graph modeling disease transmission dynamics.**

Figure 1 illustrates the random graph model we described above for a hospital with two intensive care units (ICU) and two medical/surgical units (M/S): letters on the nodes show to which set the nodes belong. The health status of the nodes, including healthy, incubation, or symptomatic, is indicated by white, pink, and red colors on the diagram. Arrows with directions show the connections, that is, $i \overset{t}{\sim} j$, among nodes within each unit. The corresponding thickness of the arrow between node i and j during the day t denotes the transmission rate, γ_{ijt}, with a heavier arrow denoting a higher transmission rate and a lighter arrow denoting a lower transmission rate. Furthermore, as shown in Figure 1, unit assignments for hospital units and thecommunity unit are dynamically changing with patient transfers and nurse clock-ins and clock-outs.

We highlight the essential role of the hidden health status in the evolution of the disease transmission network. Conceptually, the vector of health status H_t governs the formation of the disease transmission graph G_t, determining both the existence and directionality of connections between nodes. As a result, the dynamics of disease spread is inherently driven by the health status of individuals in the network. Importantly, the edge adjacency matrix χ_t and the health status vector H_t are endogenously dependent on each other, as the presence of disease transmission events (edges) directly influences the health status of the nodes and vice versa. Mathematically, hidden health status functions as a sufficient information state, encapsulating all the necessary information from the previous graphs $G_1, \ldots, G_{t-1}$. Specifically, the health status of the nurse j at time t is determined by the history of disease transmission. Nurse j remains *healthy* if and only if no infectious node has transmitted the disease to nurse j in any of the preceding periods $G_1, \ldots, G_{t-1}$, i.e.,

$$h_{j,t} = \textit{healthy} \iff x_{ijk} = 0, \quad \forall i \in \mathcal{K}, \quad k = 1, \ldots, t-1.$$

Conversely, the health status of a node directly influences the formation of connections in G_t, as only infectious nodes can transmit disease to healthy ones, thereby creating disease-transmission connections in the graph. Thus, the health status vector H_t and the adjacency matrix χ_t are intertwined through the transmission process. The vector of health statuses, H_t, serves as a comprehensive summary of past transmission events, making the entire historical sequence of graphs $G_1, \ldots, G_{t-1}$ unnecessary for determining the structure of G_t. This leads to the conditional independence of the transmission graphs given the health status:

$$P(G_t \mid G_1, \ldots, G_{t-1}) = P(G_t \mid H_t), \tag{7}$$

where H_t fully encapsulates the relevant information from prior events, and thus the previous graphs are conditionally independent given H_t.

Our dynamic network model with hidden node status contributes to the network literature by incorporating the endogeneity between the latent information and the structure of the network, that is, the presence or absence of edges between nodes affects the directions and presences of other edges. Dynamic network models with latent variables in the extant literature, e.g., latent space models and stochastic block models, assume that latent variables are drawn from exogenous random processes that are independent of the history of the edge adjacency matrix χ_t. This assumption expedites the estimation process because one can directly apply the law of total probability on hidden information to infer likelihood functions on the presence of edges; see

Kim et al. (2018) for details. However, assuming an exogenous hidden health status is unrealistic in disease transmission networks, because the presence of an edge represents that a healthy node is infected to be an infectious node. In turn, the infection determines the health status of the infected nodes for all future periods, to see this consider:

$$P(h_{jt} = healthy | H_k, d_{j,k}, s_{ijk}, i \overset{t}{\sim} j : k \leq t-1) = \begin{cases} 0 & \text{, if } \exists \quad x_{ijk} = 1 \forall i \overset{k}{\sim} j, \quad k \in 1 \dots t-1 \\ \prod_{k \in 1 \dots t-1} e^{\gamma_{jt}^T s_{jt}} & \text{, otherwise.} \end{cases} \tag{8}$$

The intuition is that if there is an infection event, the receiving node's health status cannot stay healthy. As pointed out in equation (3), the health status of nodes, in turn, impacts the edge presence probabilities; i.e.

$$P(x_{ijt} = 1 | h_{jt}, H_t, d_{it}, s_{ijt}, i \overset{t}{\sim} j) = \begin{cases} 1 - e^{\gamma_{ijt} s_{ijt}} & \text{, if } h_{jt} = healthy \\ 0 & \text{, otherwise.} \end{cases} \tag{9}$$

That is, only susceptible nurses can be infected in the disease transmission network.

4.2. Estimation Strategy

In this section, we address two key challenges in estimating the parameters θ for the hierarchical random graph model with hidden health status. A primary challenge arises from the unobservable nature of the edge adjacency matrix of the disease transmission network, $(x_{ijt})_{(i,j) \in \mathcal{K}}$. Unlike standard network models, where edges are typically observable, disease transmission events between nurses remain hidden. This is due to the nature of infectious diseases: after infection, a nurse enters an incubation period during which they are asymptomatic but still transmits the disease to other healthy nurses. Once the incubation period ends, the nurse becomes symptomatic. However, the hospital only observes this transition when the nurse calls in sick for an upcoming shift. In Section 4.2.1, we introduce a probabilistic inference framework that connects the observation model - based on when nurses report sick - to the likelihood function of our hierarchical random graph network, which models the timing of infections. Another challenge stems from the unobservable nature of connections between nodes due to the hidden health status of the nurse. Specifically, for any pair of nodes i and j, we do not observe whether $i \overset{t}{\sim} j$ or $i \overset{t}{\nsim} j$ because the health status of nodes i and j is hidden. Consequently, the interaction time between the node j and other infectious nurses nodes, S_{jt}^N, is also unobservable, as it depends on the health status of both nodes. In Section 4.2.2, we present an expectation-maximization procedure to address this challenge by handling the unobserved interaction times.

4.2.1. Likelihood Function with Latent Status. Let $Y_j \in 1 \dots T$ denote the time nurse j calls in sick, which we observe in the data. Let X_j denote the time the nurse i is infected, which we do not observe. To link infection time, X_j, to our disease transmission network, note that

$$\begin{aligned} P(X_i = \tau_i) &= P(X_{i,1} = X_{i,2} \dots X_{i,\tau_i - 1} = 0, X_{i,\tau_i} = 1) \\ &= \left(\prod_{j=1}^{\tau_i - 1} (1 - P(X_{i,j} = 1 | X_{i,j-1} = X_{i,j-2} \dots X_{i,1} = 0) \right) \cdot P(X_{i,t} = 1 | X_{i,\tau_i - 1} = X_{i,\tau_i - 2} \dots X_{i,1} = 0) \end{aligned} \tag{10}$$

Let L_i represent the time interval between the onset of symptoms of a nurse and the moment they notify the hospital by calling in sick. Although L_i remains unobservable, we only require that it be non-negative, without imposing any additional assumptions on its distribution. To see the relationship among the time periods, we have:

$$Y_i = X_i + Z_i + L_i, \tag{11}$$

that is, the absent time of a nurse equals the sum of infection time, incubation time, and time information lag. In addition, we assume that the disease incubation period Z_i is identically and independently distributed among nurses and follows a known distribution: $F_z(\cdot)$.

Our observation model consists of two types of nurses: nurses who called in sick because of infection and nurses who did not. Nurses who had not called in sick could still be infected and in the incubation period during their last shift in the data. We denote the collection of nurse features throughout the time horizon as set $\mathcal{D} = (d_{it})_{i \in \mathcal{N}, t \in 1 \dots T}$. Let τ_i denote the last time nurse i showed up in the data. For nurses who called in sick, τ_i is when nurse i called in sick, and for nurses who have not called in sick, τ_i is the last time they worked in the data. Our observation model on θ given data is specified as follows:

$$L(\theta | \mathcal{D}) = \prod_{i \in \mathcal{N}} \left(P(Y_i = \tau_i | \theta, \mathcal{D}) \right)^{\mathbb{1}_{Y_i = \tau_i}} \left(P(Y_i > \tau_i | \theta, \mathcal{D}) \right)^{\mathbb{1}_{Y_i > \tau_i}} \tag{12}$$

The first term is the probability that nurses who eventually called in sick exhibited symptoms on their last scheduled shift, while the second term is the probability that all other nurses remained at work without reporting sick.

For nurses who called in sick, let $\underline{\tau}_i$ denote the last time we saw a nurse i at work before she calling in sick. If nurse i called on sick on τ_i, nurse i must have shown symptoms between $\underline{\tau}_i$ and τ_i. If the nurse i did not call in sick, she cannot show symptoms before τ_i. Lemma 1 summarize this observation and links the unobserved infection and incubation time to the absent time:

LEMMA 1. *(a) For nurses who called in sick, we have:*

$$P(Y_i = \tau_i) = \sum_{t'=1}^{\tau_i - 1} P(X_i = t') P(\underline{\tau}_i - t' < Z_i \leq \tau_i - t'). \tag{13}$$

(b) For nurses who did not call in sick, we have:

$$\begin{aligned} P(Y_i > \tau_i) = & \left(\sum_{t'=0}^{\tau_i} P(X_i = t') \cdot P\big(Z_i \geq \tau_i - t'\big) \right) \\ & \cdot \left(1 - \prod_{t'=1}^{\tau_i} \Big(1 - P\big(X_{i,t'} = 1 \mid X_{i,t'-1} = X_{i,t'-2}, \ldots, X_{i,1} = 0\big) \Big) \right) \\ & + \prod_{t'=1}^{\tau_i} \Big(1 - P\big(X_{i,t'} = 1 \mid X_{i,t'-1} = X_{i,t'-2}, \ldots, X_{i,1} = 0\big) \Big). \end{aligned} \tag{14}$$

We show part (a) of Lemma 1 by linking the observation data, Y_i, the call in time of the nurse i, with $P(\underline{\tau}_i < X_i + Z_i \leq \tau_i)$, the probability that a nurse will show symptoms in a time interval between $\underline{t}$ and t. For nurses who did not call in sick, as in part (b) of the lemma 1, we further divide nurses into nurses who were infected but yet to show symptoms and nurses who were never infected during the horizon. Then, by the law of total probability, we could rewrite the observation model with the probability model we introduced in Section 4.1 and the known incubation distribution, $F_z(\cdot)$.

Let $\mathcal{S} = (S_{it})_{i \in \mathcal{N}, t \in 1 \ldots T}$ be the collection of the interaction times of nurses in the horizon, which is a random vector. Based on the lemma 1, we derive the likelihood function of θ conditioning on the interaction times S, by plugging the hierarchical model in equation (6) into the observation model in equation (12), resulting in the conditional likelihood function in Corollary 1:

COROLLARY 1. *$L(\theta|\mathcal{D}, \mathcal{S})$, the likelihood function conditioning on the interaction time $\mathcal{S}$ and nurse feature set $\mathcal{D}$, is given by:*

$$\begin{aligned} L(\theta|\mathcal{D}, \mathcal{S}) = & \prod_{i \in \mathcal{N}} \left(\sum_{t=0}^{\tau_i} \left((F_z(\tau_i) - F_z(\underline{\tau}_i)) \cdot \Big(\prod_{t'=1}^{t-1} (1 - \phi \circ \psi(d_{i,t'}, \theta)) \cdot \phi \circ \psi(d_{i,t}, \theta) \Big) \right) \right)^{\mathbf{1}_{Y_i = \tau_i}} \\ & \cdot \left(\left(\sum_{t=1}^{\tau_i} \left(\prod_{t'=1}^{t-1} (1 - \phi \circ \psi(d_{i,t'}, \theta)) \cdot \phi \circ \psi(d_{i,t}, \theta) \cdot (1 - F_z(\tau_i - t)) \right) \right) \cdot \left(1 - \prod_{t=1}^{\tau_i} (1 - \phi \circ \psi(d_{i,t}, \theta)) \right) \right. \\ & \left. + \prod_{t=1}^{\tau_i} (1 - \phi \circ \psi(d_{i,t}, \theta)) \right)^{\mathbf{1}_{Y_i > \tau_i}} \end{aligned} \tag{15}$$

As noted at the end of Section 4.1, the interaction vector, S, is not fully observable because when nurses become infectious or which nurses are infectious in a shift are not observable. The interaction vector, S, is a random vector whose distribution depends on the variable θ, the parameter to be estimated. In the next section, we present an expectation-maximization procedure designed to directly address this unobservable interaction vector.

4.2.2. Expectation-Maximization Algorithm for Latent Number of Nurses in Incubation. In this section, we introduce a simulation-based expectation-maximization (EM) procedure to address the unobservable interaction vector, $\mathcal{S}$, in the conditional likelihood function derived in (15). The EM procedure alternates between two steps: *expectation step*, where we estimate the distribution of the partially unobservable sample path S given the current parameter estimate $\hat{\theta}$, and *maximization step*, where we find the value of $\hat{\theta}$ that maximizes the expected likelihood function based on the simulated distribution of S.

Expectation Step. As we introduced earlier, the interaction vector, S, is a random vector whose distribution depends on the variable θ, the parameter to be estimated. Therefore, to take expectation over the vector S, we specify the relationship between $\mathcal{S}$ and θ. In the disease transmission network, the unobservable interaction vector, S, is a function of the hidden health status, $(H_t)_{t\in 1\ldots T}$, i.e., $S(H_t): H_t \to S$. H_t is a random vector whose distribution depends on the parameter θ, because the health status of nurses is driven by disease transmissions. For ease of presentation, although all stochasticity is driven by the random vector $(H_t)_{t\in 1\ldots T}$, hereafter we refer to the distribution of $\mathcal{S}$ given θ as $F_{S|\theta}(\cdot)$. Due to the complexity and high dimensionality of the random vector $\mathcal{S}$, we do not assume a parametric form for the distribution $F_{S|\theta}(\cdot)$. Instead, we propose a simulated expectation-maximization process to approximate $F_{S|\theta}(\cdot)$ using simulations and the empirical distribution conditioned on the parameter θ (Dempster et al. 1977). Let $E_{S|\theta}[L(\theta|\mathcal{D})]$ represent the expected likelihood function. The core idea of simulated expectation maximization is to iteratively identify the next parameter estimate, $\theta^{(t+1)}$, that maximizes $E_{S|\theta^t}[L(\theta^t|\mathcal{D})]$. This maximization is performed using the expectation taken over $F_{S|\theta^t}(\cdot)$, which is approximated via Monte Carlo simulations based on the current parameter estimate θ_t. Specifically, we have

$$E_{S|\theta}[L(\theta)] = \int_S L(\theta, S) dF_S(\theta). \tag{16}$$

Here the $F_{S|\theta}(\cdot)$ is not observable to us and potentially very complex to compute, therefore, we approximate this distribution by the empirical distribution generated by the simulation:

$$\hat{P}_{S|\theta}(\omega) = \frac{1}{M}\sum_{m=1}^{M} I_{\omega_m}(\omega|\theta), \tag{17}$$

where $I_{\omega_m}(\omega|\theta)$ is the indicator function for if the m^{th} sample path, ω_m, equals to ω. This approximation method is based on the Glivenko-Cantelli Theorem: As $M \to \infty$, $\hat{P}_{S|\theta}(\cdot)$ converges to the true distribution with probability 1 (Gaenssler and Stute 1993).

The Monte Carlo simulation used to derive the empirical distribution takes a trace-based approach as patient admission, movement, discharge, and nurse schedules, and corresponding features calculation directly follow the real data. In the simulation model, as we dynamically trace the nurse features and have the features to transmission rates mapping as input, we keep track of when a nurse is infected, showing symptoms, and calling in sick. Let $\Omega(\theta)$ denote the set of sample paths generated by the simulation process given parameter θ. The following proposition summarizes the expected likelihood function we wish to maximize in each iteration:

PROPOSITION 1. *Given the simulated expectation and maximization set-up, the likelihood function we want to maximize is given by:*

$$\begin{aligned}
&\hat{E}_S[L(\theta|\mathcal{D})] = \\
&\sum_{s\in\Omega(\theta)} \hat{P}_{S|\theta}(s) \prod_{i\in\mathcal{N}} \left(\sum_{t=1}^{\tau_i} \left((F_z(\tau_i) - F_z(\underline{\tau}_i)) \cdot \Big(\prod_{t'=1}^{t-1} (1-\phi\circ\psi(d_{i,t'},\theta|s)) \cdot \phi\circ\psi(d_{i,t'},\theta|s) \Big) \right) \right)^{\mathbf{1}_{Y_i=\tau_i}} \\
&\cdot \left(\left(\sum_{t=1}^{\tau_i} \prod_{t'=1}^{t-1} (1-\phi\circ\psi(d_{i,t'},\theta|s)) \cdot \phi\circ\psi(d_{i,t},\theta|s) \cdot (1-F_z(\tau_i-t)) \right) \cdot \left(1 - \prod_{t=1}^{\tau_i} (1-\phi\circ\psi(d_{i,t},\theta|s)) \right) \right. \\
&\left. + \prod_{t=1}^{\tau_i} (1-\phi\circ\psi(d_{i,t},\theta|s)) \right)^{\mathbf{1}_{Y_i>\tau_i}}.
\end{aligned} \tag{18}$$

Maximization Step. In each iteration, we update θ by maximizing the log-likelihood function:

$$\theta^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \hat{E}_S[\log L(\theta|\mathcal{D}))]$$

For the maximization step of the expectation maximization algorithm, we employ a gradient descent algorithm with initialization from multiple random starting points on the expected log-likelihood function. This approach aims to identify the solution that maximizes the log-likelihood function. Considering the complicated nature of the likelihood function, it is important to note that our proposed procedure does not guarantee the finding of parameters at the global maximum for the log-likelihood function. However, based on the results of Little and Rubin (2019), an improvement on the expected likelihood function, $E_S[L(\theta)]$, by changing θ will result in a greater improvement on the true likelihood function, $L(\theta, S)$. To address this challenge, we mitigate the risk of convergence to a local maximum by initiating expectation maximization iterations from various random starting points. This strategy helps to improve the likelihood of reaching the optimal solution.

4.3. Feature Selection via Bootstrapping of the Cox Model.

Now that we have developed the random graph network model, the next step requires identifying the best possible combination of variables for accurately predicting nurse absenteeism; i.e., selecting a set of features $\mathcal{D}$ from the total features we extracted from the dataset, as shown in Table 3. Given the computational complexity of the expectation-maximization algorithm, it is difficult to select extensive features directly from the random graph network model. Even with the application of parallel processing techniques and the utilization of high-performance computing devices, comparing the performance of a large number of feature combinations on the network model is a very lengthy process. As a solution, we offer a much more efficient bootstrapping procedure based on the Cox proportional hazard model to select the feature set to be used in the final random graph network model.

We employ a logit-based discrete-time Cox model with conditional hazard in each discrete interval to account for time-varying features. For predicting nurse absence, we consider an 'L-day' window before absence, since we do not observe the infection or symptomatic time of nurses. In other words, to predict absenteeism events, we consider features up to L days before time t: $d_{j,t-1}, d_{j,t-2} \dots d_{j,t-L}$. The value of L reflects the length of the incubation period and the information lag between a nurse showing symptoms and calling in sick. Following Cox (1972), the notion of a failure event refers to a nurse being absent due to infection, which is observable in the data. Let β denote the estimators for the Cox model and $\lambda_{i,t} = P(Y_i = t | Y_i \geq t)$ denote the conditional failure probability. The likelihood function is defined as follows:

$$\ell(\beta) = \prod_{i=1}^{N} \left[\lambda_{i,\tau_i} \prod_{t=1}^{\tau_i - 1} (1 - \lambda_{i,t}) \right]^{\mathbb{1}_{Y_i = \tau_i}} \left[\prod_{t=1}^{\tau_i} (1 - \lambda_{i,t}) \right]^{\mathbb{1}_{Y_i > \tau_i}}, \tag{19}$$

where the conditional probability $\lambda_{i,t}$ follows a logit-form:

$$\lambda_{i,t} | d_{j,t-1}, d_{j,t-2} \dots d_{j,t-L} = \frac{\exp\left(\beta_{0j} + d_{j,t-1}^{\intercal}\beta_1 + d_{j,t-2}^{\intercal}\beta_2 \dots + d_{j,t-L}^{\intercal}\beta_L\right)}{1 + \exp\left(\beta_{0j} + d_{j,t-1}^{\intercal}\beta_1 + d_{j,t-2}^{\intercal}\beta_2 \dots + d_{j,t-L}^{\intercal}\beta_L\right)}. \tag{20}$$

Following the bootstrapping procedures for the Cox model as specified by Harrell et al. (1996), we evaluate the predictive power of each feature combination by the Area under Curve (AUC) metric. In this procedure, we split the nurses into multiple training and testing sets with replacement. In each bootstrap, we fit the parameters on the training set and evaluate its AUC on the testing set. Each feature combination undergoes 500 bootstrapping evaluations and we take the average out-of-sample AUC as its performance metric.

A feature selection process (Guyon and Elisseeff 2003) typically begins with a simple comparison of statistics across the control and treatment groups as previously presented in Table 1. In Table 1, we see that the

infected nurses have more interaction with infectious patients and work in large units where the interaction with other nurses is also high. Moreover, nurses with less experience are also more likely to be infected. Therefore, our first strategy in selection is to select features that are closely related to or affecting these key differentiators. Given these initial insights on likely useful predictor variables, we next apply a backward selection strategy to locate the best feature set. Specifically, starting with the model containing all relevant features, we sequentially remove the feature terms and monitor the resulting AUC value. Additionally, for each feature combination, we monitor the VIF of each feature (variation inflation factor) to eliminate features whose VIF is greater than 10. This step is to make sure that the selected feature set does not have multicollinearity issues. Table 4 in the appendix presents the results of the final selected feature set, which achieved an AUC of 0.74.

Observe that the coefficient associated with the patient-to-nurse ratio is positive and significant. This is expected since the more patients a nurse interacts with, the busier the nurse is, hence, the more likely the nurse may be infected by a patient or by other nurses who are helping with patients. The positive and significant association between nurse absenteeism and COVID-19 interaction time is intuitive and self-explanatory— more infectious patients handled by a nurse lead to higher chances of nurse absenteeism. The experience of nurses handling COVID-19 patients, measured by their cumulative interaction time with COVID-19 patients before the shift, is negative and significant, potentially indicating nurse learning effects. That is, the nurse gains more experience after treating more COVID-19 patients and becomes more careful and better aware how to avoid transmission. As a result, the risk of getting infected decreases. Furthermore, location and time-fixed effects are also significant in predicting nurse absenteeism, which indicates variability in nurse absenteeism caused by infections across different times and hospitals.

4.4. Estimation Results and Performance

So far, we have developed the random graph network model in Sections 4.1 and 4.2 and identified a promising set of predictor variables (features) to use in the model in Section 4.3. We are now ready to apply the random graph network model with these features to a hospital dataset to estimate the transmission rates of patient-to-nurse (P-N), nurse-to-nurse (N-N) and community-to-nurse (C-N). Due to the large size of our samples and computational complexities, we demonstrate the estimation results on the largest hospital in the IU Health system during the initial wave of the COVID-19 pandemic (3/11/2020 - 7/16/2020). In this hospital, 1300 nurses managed the care of 685 COVID-19 patients during that period. During the same

period, the hospital experienced 171 cases of absenteeism of nurses due to COVID-19 infections. For the incubation period distribution, we use the results of McAloon et al. (2020), where the incubation time is best described by a log-normal distribution with parameters $\mu = 1.63$ and $\sigma = 0.5$. This translates to an average incubation period of around 5 days, with a standard deviation of approximately 2 days. In the numerical tests, we find the following parameter values to be suitable: the number of simulations to generate the empirical distribution of the nurse infection sample path in the expectation-maximization method is set to 500,000 and the number of random starting points of the expectation-maximization algorithm is set to 50.

Table 6 reports the mapping of features to transmission rates output from our expectation-maximization algorithm. In line with the Cox prediction model, we find that workload, measured by patient-to-nurse ratio, positively contributes to all three sources of transmission rates. In addition, the experience of the nurses handling COVID-19 patients consistently translates into lower transmission rates.

We demonstrate the performance of our method by comparing the simulated infection outcomes generated using parameters derived from our procedure with the actual count of infections reported in the study hospital during the same period. Figure 2 shows the performance of our estimation procedure compared to the real absence by infection counts. In Figure 2, we show counts of nurse infection events (call-in sick events) every 10 days from the beginning of the pandemic. At each time point, we compare the predicted results, three colored stacked bars in Figure 2, and the real infections, the gray bar in Figure 2. As shown in the chart, since our estimation procedure provides transmission rates from different sources, we can identify the count of infections by each source: nurse-to-nurse transmission, patient-to-nurse transmission, and community-to-nurse transmissions. In contrast to early news reports from Wong (2020) where 93% of nurse infections are believed to occur in the community, we find that only around 22% of nurse infections are in the community and 78% contribute to hospital infections.

4.5. Validation Checks

Our random graph model with hidden health status inherently requires a sophisticated structure and estimation procedure. To evaluate the identification and consistency of our estimation approach, we perform two focused validation checks: applying the parameters to data from another hospital and conducting a simulation-based parameter recovery study.

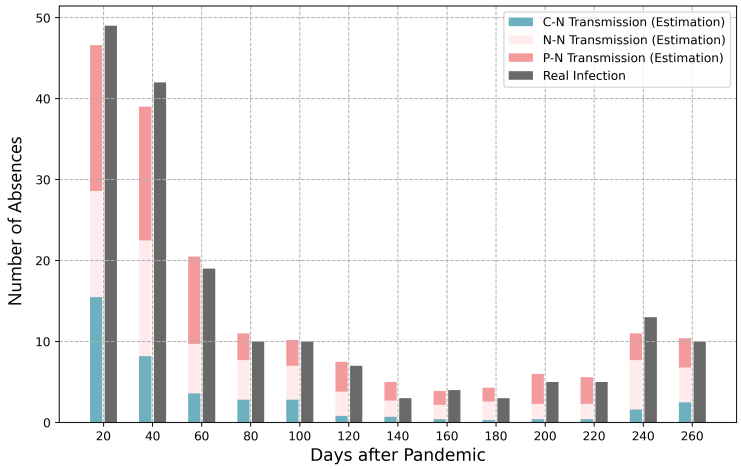


Figure 2 **Predicted Absences due to Infection via Estimated Feature to Transmission Rates Mapping Compared to the Real Infection Counts**

Out-of-Sample Validation. We first test the predictive performance of our fitted parameters on a hospital that closely resembles our original "training hospital" in size, geographical location, and patient mix. We refer to this second facility as the "test hospital." Both are large academic hospitals in the metropolitan area. Figure 3 compares the predicted nurse absences (due to infection) at the test hospital with the actual number of infected-nurse absences. As shown, parameters estimated at the training hospital also predict nurse infections robustly at the test hospital.

Parameter Recovery. Next, we conduct a simulation-based parameter recovery study following McLachlan and Krishnan (2008), to verify that our procedure accurately recovers known parameters and to address potential identification concerns. We generate synthetic data on nurse absenteeism from a "ground truth" parameter set, obtained by multiplying each element in θ by a random coefficient drawn from a log-normal distribution with parameters (μ, σ). Denote this random multiplier vector by $\Gamma_{\mu,\sigma}$. Each realization of $\Gamma_{\mu,\sigma}$ serves as a ground truth for the recovery study. For each ground truth, we produce 150 simulated nurse-absenteeism datasets, re-estimate θ on each, and compare the fitted parameters to the known values.

To determine whether the fitted parameters are statistically equivalent to the ground truth, we adopt an equivalence test using the Two One-Sided Tests (TOST) procedure (Schuirmann 1987). The null hypothesis (H_0) states that the difference between the estimated and true parameter is at least a predefined margin Δ,

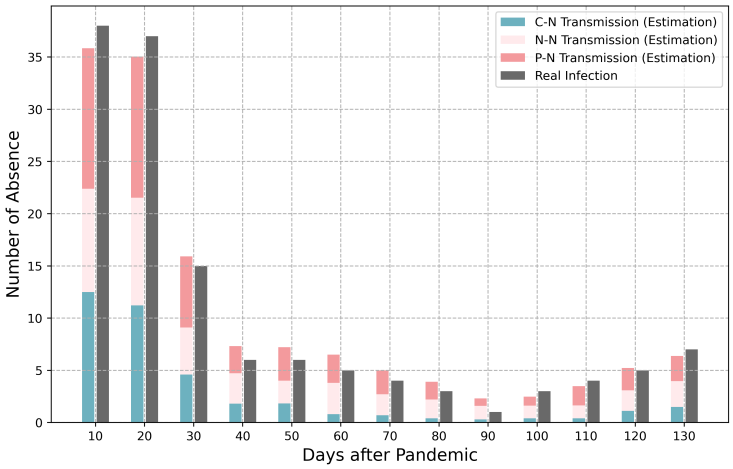


Figure 3 Prediction Results of the Fitted Parameters on Another Comparable Hospital

i.e., $|\theta_{\text{fit}} - \theta_{\text{true}}| \geq \Delta$. The p-value is computed as

$$p = P\big(T \leq T_{\text{lower}}\big) + P\big(T \geq T_{\text{upper}}\big), \tag{21}$$

where T is the test statistic, and $T_{\text{lower}}, T_{\text{upper}}$ are the critical values corresponding to the equivalence bounds. A small p-value ($p < \alpha$) allows us to reject H_0, indicating that the fitted and true parameters are statistically equivalent.

(μ, σ)	$p < 0.01$	$0.01 \leq p < 0.05$	$0.05 \leq p < 0.1$	$p > 0.1$
(0.2,0.2)	4	9	6	0
(0.3, 0.3)	3	11	4	1
(0.4,0.4)	3	9	7	0
Total Number of Parameters		19		

Table 2 Results of the Parameter Recovery Study at Different Coefficient Distributions

Table 2 reports the different levels of the results of the equivalency tests when the ground truths are generated with log-normal distributions with different parameters. The first column of Table 2 specifies the parameters of the distribution from which the multipliers are drawn to generate the ground truth. In the random graph model, we estimate a total of 19 parameters. The top row of Table 2 reports the $p-$value cutoff values for the equivalency tests between the parameter fitted by our algorithm and the ground truth. For example, when the ground truth multiplier is drawn from a lognormal distribution with parameter $(\mu, \sigma) = (0.2, 0.2)$, for 4 of 19 parameters, we can reject the null hypothesis that the fitted parameter is not

equivalent to the ground truth with $p \leq 0.01$, for 9 of 19 parameters we can reject the null hypothesis with $0.01 \leq p \leq 0.05$, and for the rest of the 6 parameters, we can reject the null hypothesis with $0.05 \leq p \leq 0.1$. That is, with $(\mu, \sigma) = (0.2, 0.2)$, we can reject the null hypothesis for all $4+9+6 = 19$ parameters for $p \leq 0.1$.

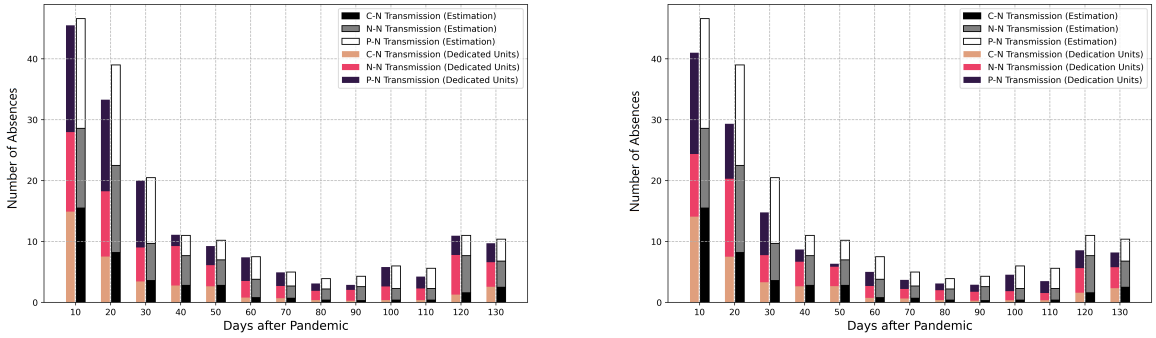
5. Counterfactual Analyses

Using real IU Health data as input for the random graph model introduced in Section 4, we evaluate the effectiveness of both staffing strategies and infection mitigation policies using counterfactual analyzes. Specifically, we implement policies such as establishing dedicated units for infectious patients, determining optimal staffing levels, assessing the effects of nurse testing at different intensities, and evaluating the impact of early vaccination at various levels of efficacy and compliance. We use trace-based simulation strategies: (1) Patient admission, transfer, and discharge information directly follows historical patient census data. (2) Nurse scheduling information also follows directly the historical workflow data. The above inputs for the model may be changed by the staffing strategies or mitigation policies; we provide details of such changes in each subsection. Similarly to the previous section, we perform counterfactual analyzes on the largest hospital in the IU Health system during the initial wave of the COVID-19 pandemic (3/11/2020 - 7/16/2020). To reiterate, during this period, 1,300 nurses in the hospital cared for 685 COVID-19 patients. Of these nurses, 171 were absent from work due to COVID-19 infection.

5.1. Dedicated Units for Infectious Patients

Designating dedicated units for infectious patients is a common strategy in hospitals to contain the spread of diseases. Typically, two levels of dedication are observed in practice: patient dedication and patient-nurse dedication. Although both levels isolate infectious patients in a single unit, patient-nurse dedication goes a step further by assigning a fixed set of nurses exclusively to infectious patients, thereby intensifying disease containment among healthcare staff.

In our dedicated units' counterfactual analyses, we allocate two units (one Medical/Surgical unit and one ICU) to exclusively admit infectious patients. To simulate this setting, we identified the hospital's Medical/Surgical unit and ICU that accommodated the highest number of COVID-19 patients and designated these as dedicated units. We then substitute the trajectories of all non-COVID-19 patients in these units with COVID-19 patients from other units, considering admissions occurring on the same day or within a maximum 2 day range.



(a) Patient dedicated units

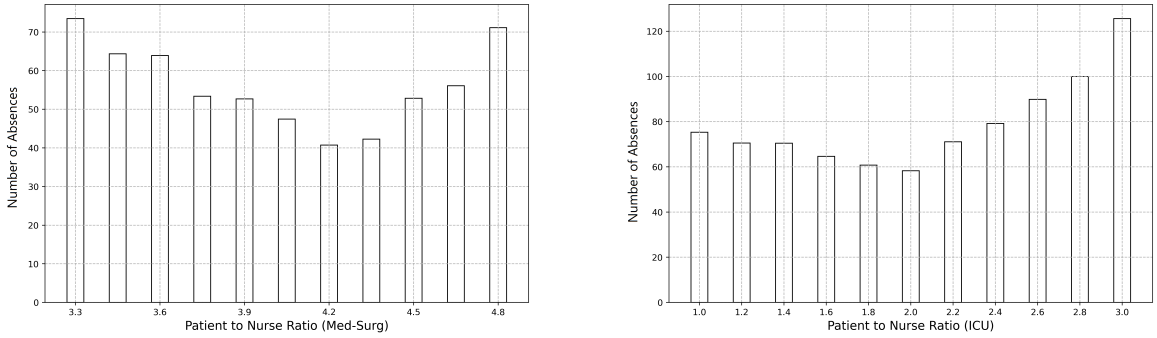
(b) Patient and nurse dedicated units

Figure 4 Total predicted absenteeism due to infections with dedicated units at different dedication levels

Figure 4 illustrates the influence of establishing dedicated units within the hospital on nurse absenteeism due to infections across the two dedication levels. Figure 4a shows the total number of nurse absences due to infection with dedicated units at the patient level alone. When compared to the estimated infection counts without dedicated units, it becomes evident that the patient dedication level offers limited efficacy in reducing the overall number of infections. On average, having units at the patient dedication level results in a modest 5% reduction in total infection counts. This reduction primarily stems from patient-to-nurse infections, possibly attributed to the increased experience of nurses handling COVID-19 patients in dedicated units. Conversely, as seen in Figure 4b, the impact of dedicated units operating at both patient and nurse dedication levels is significantly more effective in reducing total nurse absenteeism, particularly in curbing nurse-to-nurse infections. Compared to the predicted result, this higher level of dedication lowers total absenteeism by 16% and specifically nurse-to-nurse infection by 26%. This is because a higher level of dedication isolates the work of nurses in dedicated units, who face heightened infection risks, from nurses in regular units.

5.2. Staffing Level and On-Call Nurses

Staffing decisions, including determining the patient-to-nurse ratios (workload) and responses to nurse absenteeism, play an important role in hospital operations during the pandemic. As observed in Section 3, nurses with a higher workload (measured by patient-to-nurse ratio) have an increased likelihood of being absent due to infection. On the other hand, a lower workload necessitates scheduling more nurses per shift, therefore subjecting a greater number of nurses to potential infection exposure in the hospital. As a result, the hospital faces a trade-off when determining staffing levels. They must find a delicate balance: the workload should not



(a) M/S Units (b) ICU

Figure 5 Total number of absences at different workloads

be too high, where scheduled nurses face heightened risks, nor too low, where an increased number of nurses are exposed to infectious diseases. On top of this, how to respond to nurse absenteeism caused by infections raises an interesting question: What is the impact of one nurse's absence on other nurses, and what is the value of having 'on-call' nurses ready to cover the shifts of absent nurses?

In the dataset, in total of 171 nurses were infected within 131 days of the pandemic. Based on our estimation, 77 of those infections were recorded in Medical/Surgical (M/S) units and 94 absences in the ICU. To perform counterfactual analyzes, we adjust the staffing schedules for nursing shifts according to different workloads. When adjusting the workload in M/S units, we hold the workload in the ICU units constant, and vice versa. Figures 5a and 5b show the total number of absences due to infection in M/S units and ICUs, respectively, across different staffing levels. The figures show that an ideal patient-to-nurse ratio falls between 3.9 and 4.5 for M/S units and between 1.8 and 2 for ICUs. Setting the workload in M/S units and ICUs at the nurse absenteeism minimizing values of 4.2 and 2, respectively, Figure 6 shows the total number of absences due to infections at the ideal workload compared with estimation. Overall, maintaining an ideal workload leads to a reduction of approximately 25% in the total number of infections. This reduction primarily stems from decreases in nurse-to-nurse and patient-to-nurse infections. The ideal workload, higher than the workload observed in the real data during the early stages of the pandemic (the average was around 3.5 in M/S units and 1.3 in ICU units for the first two months of the pandemic), schedules fewer nurses per shift to alleviate infections. This increase of the workload lowers the likelihood of disease transmission among nurses by lowering nurse exposure to patients. As a response strategy for nurse absenteeism, we show the

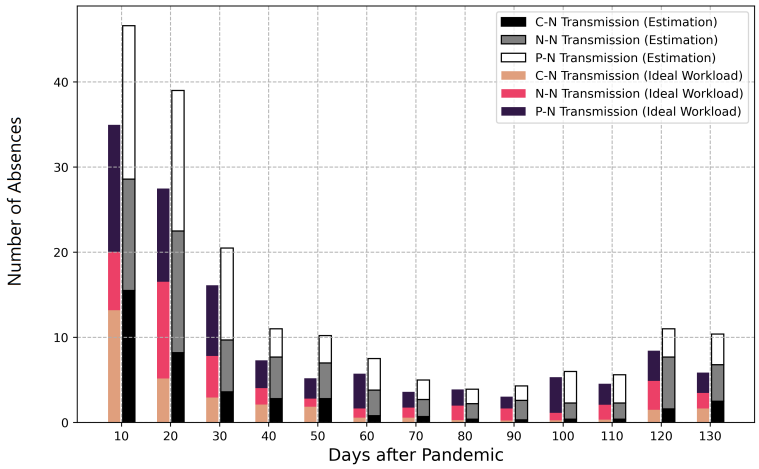


Figure 6 **Total number of absences due to infection when workload is set to be ideal**

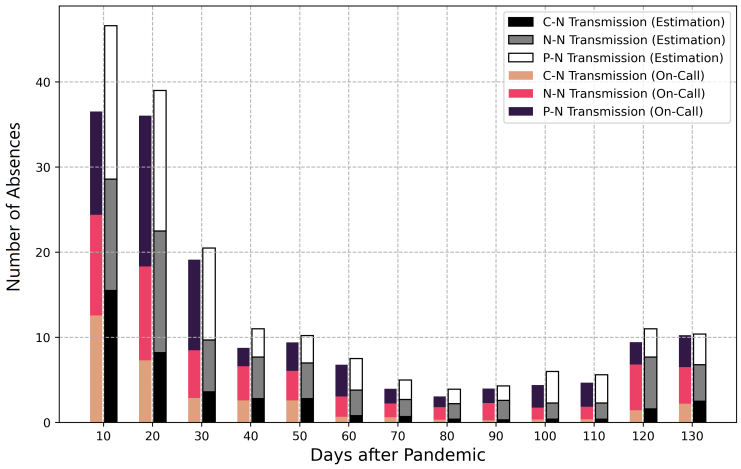


Figure 7 **Total number of absences due to infection when using on-call nurses**

value of 'on-call' nurses who cover the shifts of absent nurses. In cases of nurse absences due to infection, the hospital's current practice is to not cover the absence, potentially resulting in excessively high workloads. In Figure 7, we assume that the hospital has available on-call nurses from outside sources to cover nurse absences due to infection. This simple strategy would lead to an 8% reduction in total absences due to infection.

5.3. Pre-shift Testing for Nurse Infection

Access to testing enables hospitals to detect infected nurses during the incubation period, well before they exhibit symptoms or call in sick. This early detection allows managers to isolate those who test positive, thereby reducing the chance of further disease spread among staff. Figure 8 shows the total number of nurse absences due to infection (including both patient-to-nurse and nurse-to-nurse transmission) across a range of testing intensities. A testing intensity of 0 indicates that no nurses are tested prior to each shift, while an intensity of 1 indicates that every nurse is tested.

Even at 100% testing, we cannot completely eliminate nurse-to-nurse infection. A small possibility remains that nurses could contract infections from patients during the shift itself and transmit the virus to colleagues before any subsequent tests take place. Consequently, Figure 8 demonstrates that higher testing intensities generally curb nurse-to-nurse transmissions but do not eradicate them entirely.

In addition, we account for asymptomatic nurses, who are believed to have lower infectivity than symptomatic nurses due to reduced viral load and the absence of frequent coughing or sneezing (Sayampanathan et al. 2021). Nonetheless, removing asymptomatic nurses from the workforce after a positive test can inadvertently raise workload for the remaining staff, thereby increasing their exposure to infected patients. To model asymptomatic infections, we draw on existing research suggesting that around 30% of infected nurses are asymptomatic (Oran and Topol 2020). Specifically, for every k symptomatic infections among nurses, we assume $c \sim \text{Binomial}(k, p)$ asymptomatic infections, where $p = \frac{3}{7}$.

Because these asymptomatic but test-positive nurses must also be isolated from work, high testing intensities reduce one major source of transmission (nurse-to-nurse) while simultaneously placing greater demands on the remaining nurse workforce. Figure 8 thus illustrates this trade-off: testing reduces transmissions among coworkers but can eventually inflate patient-to-nurse transmissions if staffing capacity becomes strained. In real operations, the availability of on-call or "reserve" nurses to maintain safe staffing levels could help balance these competing effects.

5.4. Vaccination at Different Efficacy and Compliance Rates

Even though vaccination typically is not widely available at the start of an infectious disease outbreak, it stands out as one of the most effective methods to safeguard healthcare workers and eliminate the spread of the disease. In our counterfactual analysis, we investigate a realistic scenario in which combined factors of efficacy and compliance rates determine the overall effect of vaccination.

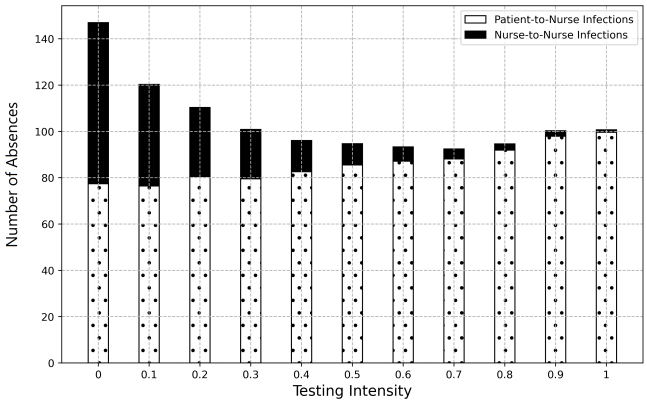


Figure 8 **Total number of absences due to infection under varying testing intensities, assuming tests detect both symptomatic and asymptomatic nurses.**

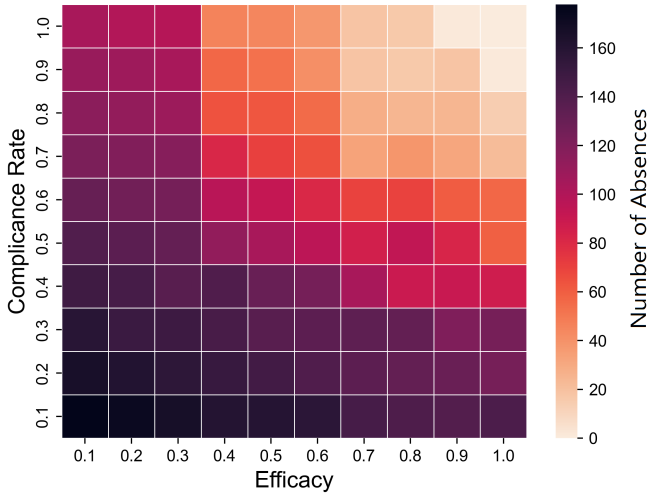


Figure 9 **Total number of absences due to infection at different efficacy and compliance rate**

Figure 9 shows a heat map that shows the total number of nurse infections at different vaccination efficacy and compliance rates. In particular, the graph indicates that the top-left corner, where the vaccination compliance rate is high but efficacy is low, exhibits fewer total infections compared to the bottom-right corner, where the vaccination compliance rate is low but efficacy is high. One explanation is that a high compliance rate may be more important, as it contributes to establishing herd immunity, thus reducing transmission rates even with vaccines of moderate efficacy (Osterholm et al. 2012).

6. Discussion

Nurses are among the most important resources in hospitals, both in terms of hospital operations and patient outcomes. Increasingly prevalent outbreaks of infectious diseases, including major and persistent outbreaks, significantly impact the well-being of nurses and disrupt hospital operations. While striving to provide essential care to patients, nurses face alarmingly high infection rates, resulting in considerable challenges to managing absenteeism due to illness, a critical concern for hospital management. Therefore, understanding the underlying causes and origins of nurse infections and insights into deploying nurses effectively based on infection awareness is of great interest to hospital managers. Using data during the first wave of the COVID-19 pandemic from IU Health, a large network of hospitals, we identify operational factors and clinical factors in predicting nurse absenteeism caused by infections. In modeling infection-aware nurse staffing, our study connects the nurse staffing model with a random graph model with hidden health status to capture the underlying disease transmission network. The granularity of our data allows us to dynamically capture the disease transmission rates for each nurse during each shift. Our study relies on numerical counterfactual analyses to provide insights for hospital managers in infection-aware nurse staffing. For future research, analytical models of the manager's decision-making process regarding nurse staffing levels, scheduling decisions, testing, mitigation policies, and patient admission and routing policies may provide additional insights for hospital managers.

Turning to the specific context of our analysis, note that our assumption that nurses do not return to work in a short time may not be valid in some environments. For example, nurses infected by the later variants of COVID-19 may come back to work in a few days, and possibly with some level of immunity to the particular variant. Thus, an understanding of the nature of the disease and the development of a model that accurately incorporates this phenomenon when estimating transmission rates would be valuable to understanding the dynamics of nurse infections.

Testing other disease mitigation policies also provides opportunities for future research. Of particular value may be social distancing among nurses as we find nurse-to-nurse infections are significantly prevalent. Also of potential value may be interventions that consider the matching of nurses who are in the incubation period of infection (not yet showing symptoms) and infectious patients as this will reduce the transmission rates of healthy nurses. Future research should investigate and compare the efficacy and cost-effectiveness of such interventions.

References

Al Maskari Z, Al Blushi A, Khamis F, Al Tai A, Al Salmi I, Al Harthi H, Al Saadi M, Al Mughairy A, Gutierrez R, Al Blushi Z (2021) Characteristics of healthcare workers infected with covid-19: A cross-sectional observational study. *International Journal of Infectious Diseases* 102:32–36.

Anand R, Gupta A, Gupta A, Wadhawan S, Bhadoria P (2012) Management of swine-flu patients in the intensive care unit: Our experience. *Journal of Anaesthesiology, Clinical Pharmacology* 28(1):51.

Baker R, Wang W (1993) Developing and testing the delay-time model. *Journal of the Operational Research Society* 44(4):361–374.

Beckley R, Weatherspoon C, Alexander M, Chandler M, Johnson A, Bhatt GS (2013) Modeling epidemics with differential equations. *Tennessee State University Internal Report* .

Blanchard P, Bolz GF, Krüger T (2005) Mathematical modelling on random graphs of the spread of sexually transmitted diseases with emphasis on the hiv infection. *Dynamics and Stochastic Processes Theory and Applications: Proceedings of a Workshop Held in Lisbon, Portugal October 24–29, 1988*, 55–75 (Springer).

Brunson JC, Laubenbacher RC (2018) Applications of network analysis to routinely collected health care data: a systematic review. *Journal of the American Medical Informatics Association* 25(2):210–221.

Burt RS (1980) Models of network structure. *Annual review of sociology* 6(1):79–141.

Carrington PJ, Scott J, Wasserman S (2005) *Models and methods in social network analysis*, volume 28 (Cambridge university press).

Chan CW, Farias VF, Bambos N, Escobar GJ (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Operations research* 60(6):1323–1341.

Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2):187–202.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1):1–22.

Drakopoulos K, Zheng F (2017) Network effects in contagion processes: Identification and control. *Columbia Business School Research Paper* (18-8).

Eames KT, Keeling MJ (2002) Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the national academy of sciences* 99(20):13330–13335.

Easton FF, Goodale JC (2005) Schedule recovery: Unplanned absences in service operations. *Decision Sciences* 36(3):459–488.

Elkholy AA, Grant R, Assiri A, Elhakim M, Malik MR, Kerkhove MDV (2019) Mers-cov infection among healthcare workers and risk factors for death: Retrospective analysis of all laboratory-confirmed cases reported to who from 2012 to 2 june 2018. *Journal of Infection and Public Health* URL https://www.sciencedirect.com/science/article/pii/S1876034119301443.

Erdős P, Rényi A (1959) On random graphs i. *Publ. math. debrecen* 6(290-297):18.

Firouzkouhi M, Abdollahimohammad A, Rezaie-Kheikhaie K, Mortazavi H, Farzi J, Masinaienezhad N, Hashemi-Bonjar Z (2022) Nurses' caring experiences in covid-19 pandemic: a systematic review of qualitative research. *Health Sciences Review* 3:100030.

Friel N, Rastelli R, Wyse J, Raftery AE (2016) Interlocking directorates in irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences* 113(24):6629–6634.

Gaenssler G, Stute W (1993) The glivenko-cantelli theorem and its extensions. *Encyclopedia of Statistical Sciences* 4:309–316.

Göttlich S, Herty M, Klar A (2005) Network models for supply chains. *Communications in Mathematical Sciences* 3(4):545–559.

Green LV, Savin S, Savva N (2013) "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* 59(10):2237–2256.

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *Journal of machine learning research* 3(Mar):1157–1182.

Harith AA, Ab Gani MH, Griffiths R, Abdul Hadi A, Abu Bakar NA, Myers J, Mahjom M, Robat RM, Zubir MZ (2022) Incidence, prevalence, and sources of covid-19 infection among healthcare workers in hospitals in malaysia. *International journal of environmental research and public health* 19(19):12485.

Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15(4):361–387, URL http://dx.doi.org/10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4.

He B, Dexter F, Macario A, Zenios S (2012) The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem. *Manufacturing & Service Operations Management* 14(1):99–114.

Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. *Journal of the american Statistical association* 97(460):1090–1098.

Jackson MR, Porter JE, Mesagno C (2023) Exploring the experiences of frontline nurses during the first 6 months of the covid-19 pandemic: An integrated literature review. *Nursing Open* 10(5):2705–2719.

Kayser A (2023) 93

Keeling MJ, Eames KT (2005) Networks and epidemic models. *Journal of the royal society interface* 2(4):295–307.

Kim B, Lee KH, Xue L, Niu X (2018) A review of dynamic network models with latent variables. *Statistics surveys* 12:105.

Kluger DM, Aizenbud Y, Jaffe A, Aizenbud L, Parisi F, Minsky-Fenick E, Kluger JM, Farhadian S, Kluger HM, Kluger Y, et al (2020) Impact of healthcare worker shift scheduling on workforce preservation during the covid-19 pandemic URL http://dx.doi.org/10.1101/2020.04.15.20061168.

Kwok KO, Tang A, Wei VW, Park WH, Yeoh EK, Riley S (2019) Epidemic models of contact tracing: systematic review of transmission studies of severe acute respiratory syndrome and middle east respiratory syndrome. *Computational and structural biotechnology journal* 17:186–194.

Lewis TG (2011) *Network science: Theory and applications* (John Wiley & Sons).

Lietz J, Westermann C, Nienhaus A, Schablon A (2016) The occupational risk of influenza a (h1n1) infection among healthcare personnel during the 2009 pandemic: A systematic review and meta-analysis of observational studies. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0162061.

Little RJ, Rubin DB (2019) *Statistical analysis with missing data*, volume 793 (John Wiley & Sons).

Liu X, Hu M, Helm JE, Lavieri MS, Skolarus TA (2018) Missed opportunities in preventing hospital readmissions: Redesigning post-discharge checkup policies. *Production and Operations Management* 27(12):2226–2250.

McAloon C, Collins Á, Hunt K, Barber A, Byrne AW, Butler F, Casey M, Griffin J, Lane E, McEvoy D, et al. (2020) Incubation period of covid-19: a rapid systematic review and meta-analysis of observational research. *BMJ open* 10(8):e039652.

McLachlan GJ, Krishnan T (2008) *The EM algorithm and extensions* (John Wiley & Sons).

Nguyen LH, Drew DA, Graham MS, Joshi AD, Guo CG, Ma W, Mehta RS, Warner ET, Sikavi DR, Lo CH, et al. (2020) Risk of covid-19 among front-line health-care workers and the general community: a prospective cohort study. *The Lancet Public Health* 5(9):e475–e483.

Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association* 96(455):1077–1087.

Oran DP, Topol EJ (2020) Prevalence of asymptomatic sars-cov-2 infection: a narrative review. *Annals of internal medicine* 173(5):362–367.

Osterholm MT, Kelley NS, Sommer A, Belongia EA (2012) Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *The Lancet infectious diseases* 12(1):36–44.

Patel M, Dennis A, Flutter C, Thornton S, D'mello O, Sherwood N (2009) Pandemic (h1n1) 2009 influenza: experience from the critical care unit. *Anaesthesia* 64(11):1241–1245.

Peck AJ, Newbern EC, Feikin DR, Issakbaeva ET, Park BJ, Fehr J, LaMonte AC, Le TP, Burger TL, Rhodes LV, et al (2004) Lack of sars transmission and u.s. sars case-patient. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322937/.

Pereira M, Williams S, Restrick L, Cullinan P, Hopkinson NS, et al. (2017) Healthcare worker influenza vaccination and sickness absence–an ecological study. *Clinical Medicine* 17(6):484.

Santos A, Cavalcante C, Wu S (2021) Discussions on the reuse of items based on the delay time modelling .

Sarkar P, Moore AW (2005) Dynamic social network analysis using latent space models. *Acm sigkdd explorations newsletter* 7(2):31–40.

Sayampanathan AA, Heng CS, Pin PH, Pang J, Leong TY, Lee VJ (2021) Infectivity of asymptomatic versus symptomatic covid-19. *Lancet (London, England)* 397(10269):93.

Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics* 15:657–680.

Scott J (2012) *What is social network analysis?* (Bloomsbury Academic).

Shakeel SM, Kumar NS, Madalli PP, Srinivasaiah R, Swamy DR (2021) Covid-19 prediction models: a systematic literature review. *Osong public health and research perspectives* 12(4):215.

Sullivan D, Sullivan V, Weatherspoon D, Frazer C (2022) Comparison of nurse burnout, before and during the covid-19 pandemic. *Nursing Clinics* 57(1):79–99.

Suwantarat N, Apisarnthanarak A (2015) Risks to healthcare workers with emerging diseases: lessons from mers-cov, ebola, sars, and avian flu. *Current opinion in infectious diseases* 28(4):349–361.

Venkatraman Y, Balas VE, Rad D, et al. (2021) Graph theory applications to comprehend epidemics spread of a disease. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience* 12(2):161–177.

Vynnycky E, White R (2010) *An introduction to infectious disease modelling* (OUP oxford).

Wang L, Hu H, Wang Y, Wu W, He P (2011) The availability model and parameters estimation method for the delay time model with imperfect maintenance at inspection. *Applied mathematical modelling* 35(6):2855–2863.

Wang W (2012) An overview of the recent advances in delay-time-based maintenance modelling. *Reliability Engineering & System Safety* 106:165–178.

Wang WY, Gupta D (2014) Nurse absenteeism and staffing strategies for hospital inpatient units. *Manufacturing & Service Operations Management* 16(3):439–454.

Whitt W (2006) Staffing a call center with uncertain arrival rate and absenteeism. *Production and operations management* 15(1):88–102.

Wong W (2020) More thanc900 mayo clinic staff in midwest diagnosed with covid-19 in past two weeks. URL https://www.nbcnews.com/news/us-news/more-900-mayo-clinic-staff-midwest-diagnosed-covid-19-past-n1248130.

Xiao J, Fang M, Chen Q, He B (2020) Sars, mers and covid-19 among healthcare workers: A narrative review. *Journal of infection and public health* 13(6):843–848.

Yang T, Chi Y, Zhu S, Gong Y, Jin R (2011) Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Machine learning* 82:157–189.

Zhao F, Wang W, Peng R (2015) Delay-time-based preventive maintenance modelling for a production plant: a case study in a steel mill. *Journal of the Operational Research Society* 66:2015–2024.

Acknowledgments

Online Supplement
Infection Aware Nurse Staffing

1. Proofs

Proof of Lemma 1

Proof: If a nurse calls in sick on shift τ_i, she must have shown symptoms between day $\underline{\tau}_i$ and τ_i. If she shows symptoms before $\underline{\tau}_i$, she would not attend shift $\underline{\tau}_i$, and if she has no symptoms until τ_i, she would not have called in sick for the day τ_i. Then, the following equation holds:

$$P(Y_i = t) = P(\underline{\tau}_i < X_i + Z_i \leq t). \tag{22}$$

By the law of total probability, we have:

$$P(\underline{\tau}_i < X_i + Z_i \leq \tau_i) = \sum_{k=1}^{\tau_i - 1} P(X_i = k, \underline{\tau}_i - k < Z_i \leq \tau_i - k). \tag{23}$$

By the independence assumption:

$$\sum_{k=1}^{\tau_i - 1} P(X_i = k, \underline{\tau}_i - k < Z_i \leq \tau_i - k) = \sum_{k=1}^{\tau_i - 1} P(X_i = k) P(\underline{\tau}_i - k < Z_i \leq \tau_i - k). \tag{24}$$

For nurse i who did not call in sick, i.e. $Y_i > \tau_i$, we have two sub-cases: (1) nurse i been infected in the horizon, but has not shown symptoms, and (2) the nurse has not been infected. We can write out the likelihood functions of a nurse did not call in sick using conditional probability in the following two terms:

$$\begin{aligned} P(Y_i > \tau_i) &= P(X_i + Z_i > \tau_i) \\ &= P(X_i + Z_i > \tau_i | X_i \leq \tau_i) \cdot P(X_i \leq \tau_i) + P(X_i + Z_i > \tau_i | X_i > \tau_i) \cdot P(X_i > \tau_i) \end{aligned} \tag{25}$$

The first term corresponds to the nurses of the first sub-case, who are infected but did not show symptoms before τ_i, and the second term corresponds to nurses who are never infected. For the second case, since the incubation period is a positive number, nurses who are never infected will have a probability 1 of not calling-in sick, i.e. $P(X_i + Z_i > \tau_i | X_i > \tau_i) = 1$, then by total law of probability, we have

$$\begin{aligned} P(Y_i > \tau_i) &= P(X_i + Z_i > \tau_i) \\ &= P(X_i + Z_i > \tau_i | X_i \leq \tau_i) \cdot P(X_i \leq \tau_i) + P(X_i + Z_i > \tau_i | X_i > \tau_i) \cdot P(X_i > \tau_i) \\ &= \Big(\sum_{t=1}^{\tau_i} P(X_i = t) \cdot P(Z_i > \tau_i - t) \Big) \cdot \Big(1 - \prod_{t=1}^{\tau_i} \big(1 - P(X_{i,t} = 1 | X_{i,t-1} = X_{i,t-2} \dots X_{i,1} = 0) \big) \Big) \\ &+ 1 \cdot \Big(\prod_{t=1}^{\tau_i} \big(1 - P(X_{i,t} = 1 | X_{i,t-1} = X_{i,t-2} \dots X_{i,1} = 0) \big) \Big) \end{aligned} \tag{26}$$

□

Proof of Corollary 1

Proof: The first term with the product is the likelihood function for nurses who called in sick. To see this, plug-in equation (6) into the corresponding term on observation model, i.e. equation specified in (a):

$$\begin{aligned} P(Y_i = \tau_i) &= P(\underline{\tau}_i < X_i + Z_i < \tau_i) \\ &= \sum_{k=1}^{t-1} P(X_i = k) P(\underline{t} - k < Z_i \leq t - k) \\ &= \sum_{t=1}^{\tau_i} \left(P(\underline{\tau}_i - t < Z_i \leq \tau_i - t | X_i = t) \cdot \left(\prod_{t'=1}^{t-1} (1 - \phi \circ \psi(d_{i,t'}, \theta)) \right) \cdot \phi \circ \psi(d_{i,t}, \theta) \right). \end{aligned} \tag{27}$$

Using the same logic, we plug in the hierarchical model in equation (6) to the term on observation model which corresponds to nurses who never called in sick:

$$\begin{aligned} P(Y_i > \tau_i) &= P(X_i + Z_i > \tau_i) \\ &= P(X_i + Z_i > \tau_i | X_i \leq \tau_i) \cdot P(X_i \leq \tau_i) + P(X_i + Z_i > \tau_i | X_i > \tau_i) \cdot P(X_i > \tau_i) \\ &P(X_i + Z_i > \tau_i | X_i \leq \tau_i) \cdot P(X_i \leq \tau_i) + P(X_i + Z_i > \tau_i | X_i > \tau_i) \cdot P(X_i > \tau_i) \\ &= \left(\sum_{t=0}^{\tau_i} P(X_i = t) \cdot P(Z_i > \tau_i - t) \right) \cdot \left(1 - \prod_{t=1}^{\tau_i} \left(1 - P(X_{i,t} = 1 | X_{i,t-1} = X_{i,t-2} \ldots X_{i,1} = 0) \right) \right) \\ &+ 1 \cdot \left(\prod_{t=1}^{\tau_i} \left(1 - P(X_{i,t} = 1 | X_{i,t-1} = X_{i,t-2} \ldots X_{i,1} = 0) \right) \right) \\ &= \left(\sum_{t=0}^{\tau_i} \prod_{t'=1}^{t-1} (1 - \phi \circ \psi(d_{i,t'}, \theta)) \right) \cdot \phi \circ \psi(d_{i,t}, \theta) \cdot P(Z_i > \tau_i - t) \Big) \cdot \left(1 - \prod_{t=1}^{\tau_i} (1 - \phi \circ \psi(d_{i,t}, \theta)) \right) \\ &+ \prod_{t=1}^{\tau_i} (1 - \phi \circ \psi(d_{i,t}, \theta)) \end{aligned} \tag{28}$$

After combining the terms, we have the desired results in proposition 1. □

Proof of Proposition 1

Proof: By combining equation (17) and (15), we can get the likelihood function as shown. □

2. Algorithms

Algorithm 1: Bootstrapping of the Cox Model

for *each re-sample* **do**

Split nurses into the training set and testing set by the 70/30 rule;

for *each specification of the model* **do**

Fit the model with the train data;

Record the coefficient estimation;

Calculate the VIF of the coefficients;

Predict outcomes using the test data;

Measure the internal AUC using the train data;

Measure the external AUC using the test data;

end for

end for

Take the average of the external AUC, Coefficient, and VIF for each model over all re-sampling of the data;

Record the average difference between internal and external AUCs. ;

Calculate the coefficient variation of the coefficients by $\frac{\sigma_{coef}}{\mu_{coef}}$;

3. Feature Definition and Feature Selection Results

Shift Specific Factors	
Variable	Definition
Ratio(both types of patients)	Average Patient-to-nurse ratio during the shift
NonCOVID Interaction	Patients × Hours spent with Non-COVID-19 patients
NonCOVID patients	Distinct number of NonCOVID Patients encountered in the same unit during shift
COVID Interaction	Patients × Hours spent with COVID-19 patients
COVID patients	Distinct number of COVID Patients encountered in the same unit during shift
Nurse Interaction	Nurses × Hours spent with other nurses in the unit
Nurses	Distinct number of Nurses encountered in the same unit during the shift
Lead Nurse	Indicator for being the lead nurse for that shift
Off_shift_time	Total off-shift time within the prediction window
Individual-Specific Factors	
Variable	Definition
C_COVID_Interaction	Cumulative Patients × Hours spent with COVID-19 patients before the shift
C_Shifts	Cumulative number of shifts attended
C_Nurses	Cumulative nurse interaction time before the shift
Fixed Effects	
Variable	Definition
Month	Month fixed effect
Location	Facility fixed effects
Unit Type	Type of the unit the nurse works

Table 3 Summary of features used in prediction.

	Estimation Statistics			
Performance	Mean	SD	95%CI	
AUC	0.74	0.12	(0.694, 0.761)	
Predictors	Coef Estimation	SD	95%CI	$P > \|t\|$
Shift Specific Features (β)				
ratio	0.084	0.042	(0.003, 0.165)	0.012***
ratio_lag1	0.042	0.064	(-0.086, 0.159)	0.075*
ratio_lag2	0.052	0.048	(-0.045, 0.131)	0.096*
Covid_interaction	0.04	0.04	(-0.007, 0.103)	0.048**
Covid_interaction_lag1	0.03	0.043	(-0.065, 0.096)	0.13
Covid_interaction_lag2	0.029	0.039	(-0.058, 0.082)	0.23
Nurses	0.052	0.04	(-0.017, 0.128)	0.09*
Nurses_lag1	0.067	0.053	(-0.047, 0.155)	0.17
Nurses_lag2	0.004	0.043	(-0.08, 0.08)	0.65
Off-shift Feature				
off_shift_time	0.071	0.08	(0.082, 0.215)	0.03**
Individual-Specific Features				
C_Covid_interaction_before_horizon	-0.068	0.072	(-0.196, -0.054)	0.008***
Fixed Effects				
month_3	0.337	0.071	(0.208, 0.498)	0.001***
month_4	-0.006	0.071	(-0.145, 0.151)	0.67
month_5	-0.274	0.073	(-0.412, -0.117)	0.001***
month_6	-0.35	0.077	(-0.487, -0.191)	0.001**
level_ICU	-0.122	0.048	(-0.218, -0.045)	0.001***
Location_1	0.018	0.1	(-0.129, 0.194)	0.56
Location_2	-0.128	0.114	(-0.31, 0.066)	0.09*
Location_3	0.161	0.11	(0.011, 0.342)	0.001***
Location_4	0.186	0.166	(-0.046, 0.443)	0.06*
Location_5	0.075	0.126	(-0.129, 0.278)	0.78
Location_6	0.093	0.094	(-0.06, 0.249)	0.17
Intercept	-5.87	0.032	(-5.936, -5.811)	0.000***
Observations				100654

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4 Feature selection results: estimated coefficients for main covariates and their 95% confidence interval (CI) and p-values.

4. Notation Table

Table 5 Model Notation

	Data
$\mathcal{D}$	Whole dataset containing features for each nurse i and each shift t
$d_{i,t}$	Features of nurse i during shift on day t
$d^c_{i,t}$	Features of nurse i during shift on day t that are related to community-to-nurse infection
$d^p_{i,t}$	Features of nurse i during shift on day t that are related to patient-to-nurse infection
$d^n_{i,t}$	Features of nurse i during shift on day t that are related to nurse-to-nurse infection
$S^n_{i,t}$	Interaction between nurse i and infected nurses in same unit during shift on day t
$S^p_{i,t}$	Interaction between nurse i and infectious patients in same unit during shift on day t
t'_i	Off-shift time of nurse i on day t
τ_i	Last shift of nurse i, call-in sick time or last shift time in data
$\underline{\tau}_i$	Time of the shift before last shift of nurse i
$t(i)$	Collection of days in which nurse i appeared in data
$\mathcal{N}$	Collection of Nurses
	Parameters
$\theta : \theta_1, \theta_2$	Parameters that maps nurse-shift-specific features to nurse-shift-specific transmission rates
$\theta_1 \in \mathbb{R}^{\lvert d^c_{i,t}\rvert + \lvert d^p_{i,t}\rvert + \lvert d^n_{i,t}\rvert}$	Coefficients of nurse-shift-specific features to nurse-shift-specific transmission rates
$\theta_2 \in \mathbb{R}$	Coefficients of nurse-shift-specific features to nurse-shift-specific transmission rates
$\gamma_{i,t}$	Nurse i 's transmission rates during shift on day t
$\gamma^c_{i,t}$	community-to-nurse transmission rates for nurse i during shift on day t
$\gamma^p_{i,t}$	patient-to-nurse transmission rates for nurse i during shift on day t
$\gamma^n_{i,t}$	nurse-to-nurse transmission rates for nurse i during shift on day t
	Variables and Functions
Y_i	Call-in sick time of nurse i
X_i	Infection time of nurse i
$X_{i,t}$	Event of nurse i infected on day t, 1 means true, 0 means false
Z_i	Incubation length of nurse i
$F_z(\cdot) : \mathbb{R}+ \to [0,1]$	Incubation length distribution

5. Random Graph Estimation Results

Predictors	Parameter Estimation ($\hat{\theta}_1$)		
	Patient-to-Nurse (γ_{pn})	Nurse-to-Nurse (γ_{pn})	Community-to-Nurse (γ_{pn})
Ratio	0.167	0.782	-
C_Covid-interaction	-0.0223	-0.041	-0.0213
Level_ICU	0.56	0.18	-
Month_3	1.32	0.90	0.23
Month_4	0.16	0.36	0.11
Month_5	-0.06	-0.24	0.1
Intercept (θ_2)	-2.364	-2.678	-3.14

Table 6 Features to transmission rates mapping on the largest hospital in IU Health Network