Submitted to *INFORMS Journal on Applied Analytics*
manuscript 0000

Delta Coverage: The Analytics Journey to Implement a Novel Nurse Deployment Program

Jonathan E. Helm* Pengyi Shi*
Indiana University Kelley School of Business Purdue University Daniels School of Business
Mary Drewes, Jacob Cecil
Indiana University Health

Amidst critical levels of nurse shortages, we partnered with Indiana University Health (IUH) System to pioneer a novel suite of advanced data and decision analytics for a groundbreaking internal travel nursing program. This state-wide program leverages a flexible pool of resource nurses who can move between the 16 IUH hospitals located in five diverse regions and serving more than 1.4 million residents. This program breaks the mold of traditional resource nurse by moving nurses *between hospitals* to dynamically respond to short-term patient census fluctuations in days rather than weeks. This paradigm shift necessitated the development of new operational protocols and analytics to execute them, including a creating two-week advance on-call list for travel and a 24-48 hour call-in decision. Our co-developed Delta Coverage Analytics Suite, launched in Oct 2021 as a Microsoft Power BI application, provides an integrated solution to support this groundbreaking initiative at an unprecedented state-wide scale, in contrast to existing nurse scheduling tools that primarily cater to single hospitals or units. The suite incorporates (i) a patient census forecast based on a deep generative model capturing complex spatial-temporal correlations and avoiding error accumulation common in traditional time-series models, which seamlessly integrates with (ii) a stochastic optimization that prescribes optimal on-call and call-in decisions. The pilot, conducted from May to June 2023, produced a remarkable 13% reduction in understaffing, with estimated annual savings of $400K and over 250 fewer understaffed shifts, all by efficiently managing the movement of only 10 nurses. As the first known program of its kind, our efforts establish new benchmarks for evidence-based and data-driven nurse workforce management, potentially transforming how healthcare institutions approach staffing challenges nationwide.

Key words: Internal travel nurse, time-series forecast, predictive-prescriptive integration

---

The decades long nurse shortage crisis has elevated to the level of global health emergency, with the United States projected to face a deficit of half a million nurses within the next two years and annual burnout and turnover rates exceeding 20%. The accelerating shortage of nurses combined with large spikes in demand has prompted hospitals and health systems to explore innovative solutions for both the short and long term. This paper presents one such such breakthrough innovation, co-developed and successfully implemented in partnership with Indiana University Health (IUH) – the **Delta Coverage (DC)** internal travel nursing program. IUH, the largest healthcare system in

* These two authors co-led the project and made equal contributions to this work.

Indiana, includes 16 hospitals and over 9,000 nurses, serving 1.4 million residents across five diverse regions spanning 14,000 square miles. The DC program, to our best knowledge, is the first *implemented, state-wide program* that utilizes a flexible pool of resource nurses, effectively moving them between the 16 IUH hospitals, to addresses understaffing challenges by harnessing the expansive reach of such large hospital system. In contrast to typical travel nursing (12-week contracts), DC employs short-term deployments, dynamically responding to geographic and temporal fluctuations in hospital occupancies. The implemented DC network design is shown in Figure 1, along with IUH's catchment area to highlight its state-wide coverage.

Our collaborative efforts led to the development of the innovative **Delta Coverage Analytics Suite**, a comprehensive solution and a pioneering implementation that leverages state-of-the-art predictive and prescriptive analytics to support the DC program, dynamically optimizing nurse deployment and staffing on an unprecedented scale. This contrasts off-the-shelf nurse scheduling analytics, which usually target individual units or hospitals, or other existing hospital analytics that prioritize physicians and patients. The distinctive dynamics and complexities of real-time nurse deployment over a large network make it difficult for existing solutions to gain traction and establish a strong foothold, leaving a gap in the market for innovations like our DC Analytics Suite to step in and pioneer a breakthrough.

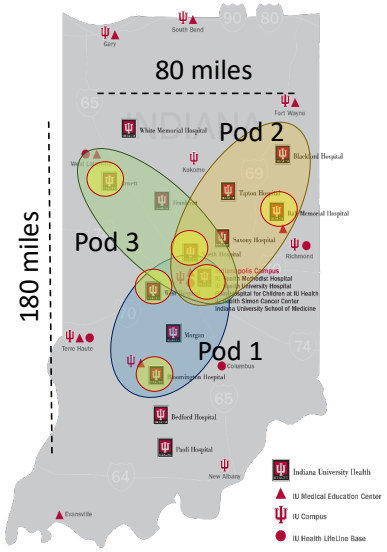


Figure 1 **DC network design. Red squares are IUH hospitals. Yellow circles indicate pilot hospitals. Colored ellipses are DC pods. DC nurses can be deployed to any hospital within their pod.**

Reduction	Understaff	Overstaff
Annual Project Shifts	250	290
Percent	13%	5%

(a) System-wide value

	Work Variety (Gini)	Sched Stability (CV)	Hospital % DC Shifts Used
Average	0.36	$0.3^{\dagger}$	19%
Equity	$0.3^{\dagger}$	0.31	$0.29^{\dagger}$

(b) Average value and equity score across all DC nurses / hospitals

Table 1 **Performance of the Delta Coverage pilot May-Jun 2023. "Work variety" is measured via Gini coefficient; smaller value means shifts are more equally divided among hospitals. "Schedule stability" is measured via coefficient of variation; smaller value is better with ≤ 0.5 (†) being very stable. "Equity" is measured by the Gini coefficient of inequality; ≤ 0.3 (†) generally considered very equitable. More details in Appendix E.**

Implementation and Impact. Launched in October 2021, the analytics suite underwent three phases of implementation. During the final pilot phase, which ran from May to June 2023, we achieved remarkable results: a **13% reduction** in **understaffing**, a **5% reduction** in **overstaffing**, and an allocation that is **fair** to participating hospitals and nurses. Table 1 provides a summary of the pilot's impact projected to annual estimates, alongside equity analysis for nurses and hospitals. All this was made possible by moving only 10 DC nurses among 6 hospitals participating in the 6-week pilot. The results demonstrate that each DC nurse is equivalent to 1.25 non-DC nurses, effectively mitigating staffing deficits. Extrapolating this impact to the United States' 1.7 million hospital registered nurses (source: Bureau of Labor Statistics) indicates the potential to almost cover the nation's half a million nurse shortfall.

The concept behind Delta Coverage is to allow highly skilled nurses to float and work on any unit, including floating to other hospitals in the network. The ultimate goal is to respond rapidly to fluctuations in staff and occupancy across the 16 hospitals. Unlike programs for traditional resource nurses, who usually float between units within a hospital and receive their assignments less than 24 hours before a shift, Delta Coverage requires sophisticated advanced planning that utilizes (i) predictive analytics to forecast occupancies for all 16 hospitals and (ii) prescriptive analytics to determine optimal on-call and call-in decisions for DC nurse transfers. To meet this urgent need, our team developed a first-of-its-kind analytics suite, seamlessly integrating state-of-the-art machine learning-based time series predictions for component (i) and stochastic optimization for (ii). Figure 2 provides a close-up of the decision support for the two stages of decisions, with the "Plan" (right panel) indicating how many nurses should be put on-call to travel 1-2 weeks in advance (e.g., from ISR to AHC in 10 days), and the "Execution" (left panel) showing how many nurses should be called in for travel 24-48 hours in advance (e.g., from Methodist Hospital to University Hospital tomorrow).

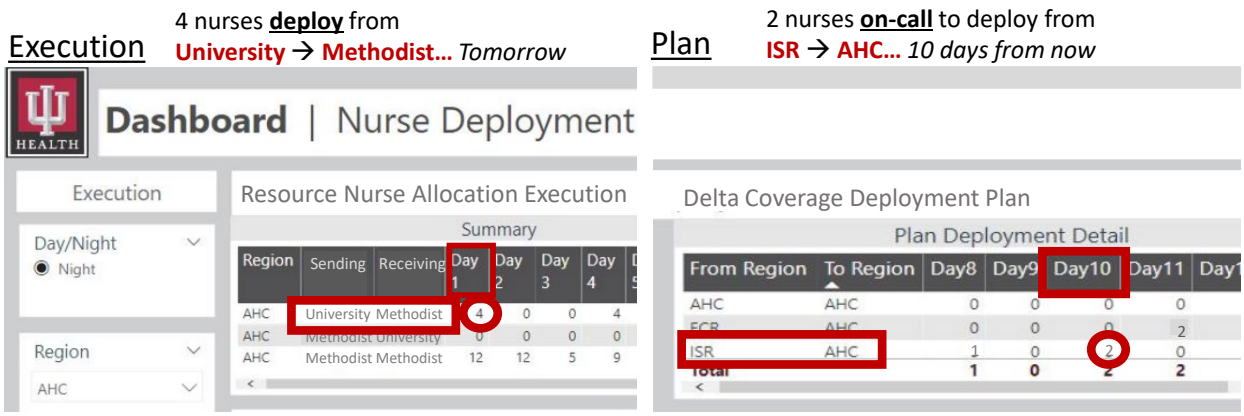
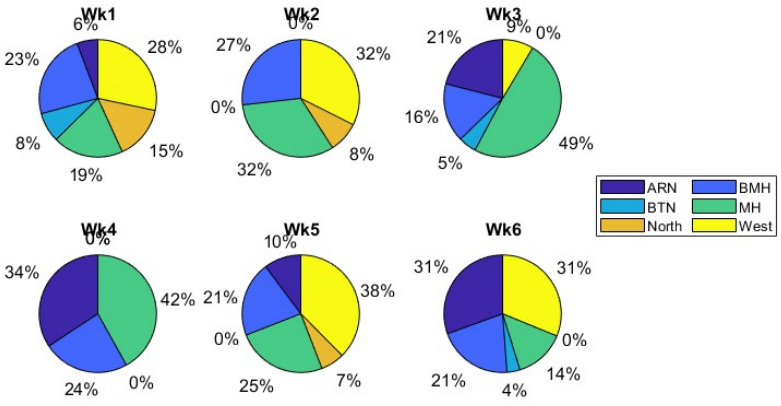


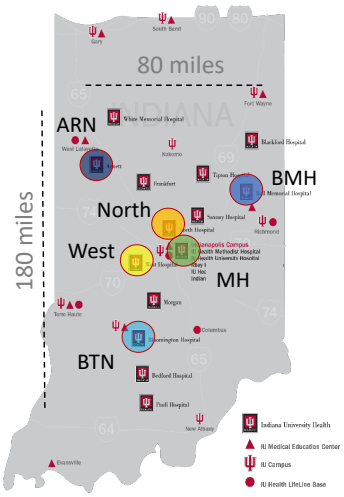
Figure 2 **Closeup snapshot of the DC Dashboard decision support for on-call and call-in decisions.**

Beyond Reducing Shortage. The reduction in understaffing achieved through our DC program has long-term societal benefits, including improved patient care, increased professional satisfaction among bedside nurses, and ultimately, lives saved (Aiken et al. 2014, Blegen et al. 2011). The long-term impact of broader deployment of our DC program on the nursing crisis is significant, given that our novel system directly addresses the primary cause of the nursing crisis - nurses leaving the profession due the pervasive issue of understaffing (Flinkman et al. 2010).

The pilot also demonstrates the desirable "fairness" feature of our DC analytics suite, benefiting both the DC nurses and participating hospitals, as evidenced by the "Equity" row in Table 1. This crucial aspect ensures the sustainability and wider adoption of the program, making it also applicable to other hospitals facing similar challenges nationwide. In particular, one significant concern voiced by Chief Nursing Officers of individual hospitals was that the urban hospitals may potentially be allocated most or all of the DC nurses, taking resources away from more rural hospitals without giving back. However, the implementation shows promising results for the hospitals located in more rural communities. Figure 3a provides a visual representation of the distribution of Delta Coverage resources among participating hospitals. The figure illustrates that, despite week-to-week fluctuations, the decisions made by the optimization engine and implemented by the DC manager result in a fair and equitable allocation of DC nurses across the participating hospitals, notably benefiting ARN and BMH, the two most rural hospitals in the pilot. See Appendix E for a comprehensive analysis of the pilot program's performance.



(a) Fraction of DC shifts allocated to each hospital by week weighted by hospital size.



(b) DC hospital map.

Figure 3 Hospital equity by week of the pilot and Delta Coverage (DC) pilot map

More broadly, the success of Delta Coverage highlights the viability and benefits of internal travel nurse programs as a solution for managing nurse shortages and optimizing workforce allocation.

This sets new benchmarks for efficiency and adaptability in addressing nurse shortages and fluctuating patient demands with data-driven and analytics-based decision-making, encouraging other hospitals to adopt similar strategies to meet the urgent demands of the healthcare landscape.

To summarize, the impact of our Delta Coverage program goes beyond immediate staffing shortage reduction. It has the potential for far-reaching impact by effectively addressing the pressing challenges of nurse shortage and burnout. In the long run, this approach promotes workforce stability and a supportive environment, resulting in a more resilient and satisfied nursing workforce. More importantly, our analysis shows that the DC program's benefits extend to rural and marginalized areas that often bear the brunt of nursing shortages (as rural hospitals face more challenges in attracting and retaining nurses due to their remote locations), disproportionately affecting access to quality healthcare and population health outcomes in these areas. By distributing DC shifts fairly among participating hospitals, we ensure rural hospitals receive the support needed to provide uninterrupted care to their communities, effectively enhancing treatment accessibility in underserved regions. The success of the program in promoting both workforce stability and equitable distribution of nurses exemplifies the transformative power of analytics-based OR solutions.

Paper Organization. In the remainder of this paper, we detail our three-year journey from the development and implementation of a necessity-driven innovation to a sustainable data-driven approach that has the potential to turn the tide against the nursing shortage crisis. In Section 1, we present an overview of our Decision Support System (DSS), the Delta Coverage Analytics Suite, and outline the technical and practical challenges encountered, underscoring our main contributions that lay the groundwork for subsequent sections:

- To overcome the technical challenges, we first describe the novel, multi-hospital and multi-unit nursing demand forecast based on a deep generative model in Section 2. We then introduce in Section 3 the prescriptive framework based on the stochastic optimization. In Section 4 we discuss the *seamless integration* of forecast and optimization: the generative model structure perfectly complements our quasi-Monte Carlo approach to overcome the curse of dimensionality in our large-scale decision optimization, which is critical because it must be solved daily even with limited computational resources.
- In Section 5 we discuss the journey to launch the pilot implementation including practical challenges and our tiered implementation approach to build trust for deploying OR analytics for operational improvement. We conclude this paper with ongoing work in Section 6.

1. Delta Coverage Analytics Suite Details and Challenges

Our analytics suite was implemented in October 2021 as a Microsoft Power BI application and went through three phases: (i) live testing from October 2021 to April 2022, (ii) program re-design

and refinement with the leadership team, and (iii) pilot with end-user adoption from May 2023 to June 2023. The implemented analytics suite is fully integrated with IUH's data-warehouse and staffing data systems, and the following procedures run on a daily or weekly basis:

1. On Monday, based on the demand forecast, scheduled nurses at each hospital, and available Delta Coverage resource nurses, the model determines the on-call list for a one-week period two weeks in advance (lookahead for 21 days).
2. Each day at 4am, update the patient census data and forecasts and determine actual deployment decisions for the following day.
3. Output is loaded into the Microsoft PowerBI dashboard to support decision making. The result of the previous day's actions (deployment, census, and updated census prediction) are recorded for program evaluation and control charting to monitor ongoing system accuracy.

Figure 4 provides a schematic of the DC Analytics Suite design; see Appendix D for detailed explanation.

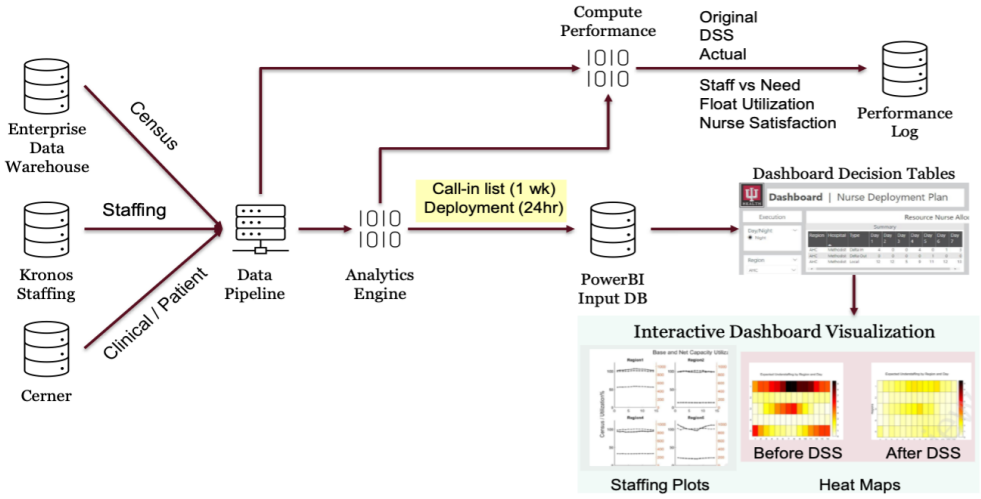


Figure 4 Delta Coverage Decision Support Data and Workflow.

1.1. Challenges

Given the goal of providing 1-2 weeks notice to nurses who will be put on-call to travel and 24-48 hour notice as to whether a nurse will be called-in, decisions must be made without full information surrounding nursing supply and demand. This required both accurate nurse demand forecasts across the 16 hospitals over multiple days as well as dynamic decisions that consider complex spatial-temporal demand correlations while accommodating nurse preference and availability. However, these models come with significant technical challenges due to hard-to-predict occupancy fluctuations and multiple shift rotations that introduce additional correlations, influencing the decisions throughout the network.

The primary challenge lies in capturing the complicated spatial-temporal correlations in patient census at different hospitals over the next 21 days. As a most obvious example of why correlation is important, consider an infectious disease outbreak, where the underlying disease spread drives hospitalizations over different regions. Even without a major public health event, weather, hospital diversions, and patient transfers among units/hospitals create complex non-linear correlation between hospitals. In our case, the decision structure that transfers nurses between hospitals complicates the system further, which contrasts to typical nurse staffing with newsvendor-type models, because (1) traveling to remote hospitals requires deployed nurses to stay there for multiple days ("secondment"), which makes decision critically depend on correlated census patterns over multiple days; (2) the DC nurse pool is shared across 16 hospitals forcing the decision framework to also account for spatial correlations. Hence, nurse staffing in such a large-scale hospital network requires accounting for spatial-temporal correlations from both the predictive and prescriptive components.

Beyond the technical challenges, we also face numerous practical obstacles. Penetrating the nursing industry with OR analytics has been exceptionally challenging due to several factors. First, the nursing profession relies heavily on established practices and protocols, leading to a resistance to change and potential hurdles in adopting novel technologies and decision analytics. Convincing the industry to embrace a groundbreaking solution requires substantial evidence of its efficacy and benefits. Second, implementing large-scale decision analytics in the nursing field has been limited by the absence of comprehensive solutions tailored specifically to this industry. Off-the-shelf nurse scheduling analytics usually target individual units or hospitals, while other hospital analytics prioritize physicians and patients, often overlooking the distinctive dynamics and complexities of nursing. These challenges make it difficult for analytics solutions to gain trust to establish a strong foothold. We further discuss these challenges during the implementation in Section 5.

1.2. Literature Review and Main Contributions

We review two main streams of literature that relate to the predictive and prescriptive components of our work.

Time-series Forecast. Traditional time-series forecast tools like autoregressive models (AR and ARIMA) or queueing-based simulations rely on parametric assumptions, such as linear dependence or Poisson arrival processes. However, these models lack flexibility in handling highly time-varying dynamics and complex nonlinear correlations. On the other hand, typical machine-learning prediction models often provide point estimates rather than the needed *distribution* for decision making under uncertainties. Recent advancements in generative models, Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) have the advantages of providing distributions as the output. Time-series generative models use GAN or VAE and combined with RNN, e.g., see

Mogren (2016), Esteban et al. (2017), Desai et al. (2021), among others. TimeGAN (Yoon et al. 2019), considered as the current state-of-art method, combine autoregressive models with GANs and aligned the latent representations of real and generated data. However, these generative models have one primary limitation: learning step-wise conditional distributions that may accumulate errors and overlook key temporal patterns essential for downstream tasks; see more discussion in Section 2. Moreover, they often lack theoretical justification, interpretability, and fail to consider the structural insights of realistic problems. In contrast, the predictive model we developed in this work effectively addresses the error accumulation issue and is domain adapted.

Nurse Staffing and Deployment. Nurse scheduling is a topic that has been well studied in the OR/MS literature, e.g., see Griffiths et al. (2020), Saville et al. (2019) for a comprehensive review. Recent advance on analytics have helped incorporating predictive analytics into the nurse scheduling, e.g., see, Ban and Rudin (2019), Anderson et al. (2022), Spetz (2021), Shi et al. (2023), Zlotnik et al. (2015), among others. These studies emphasize the significant impact that sophisticated prediction models can have on optimizing nurse staffing levels and improving patient outcomes. The most relevant paper to our work is Hu et al. (2021), who use predicted patient demand to guide base and surge nurse staffing in ED. It is important to highlight that this stream of literature has predominantly focused on staffing within individual or hospital units, which operates on a much smaller scale compared to our work. Consequently, these studies usually do not consider complex spatial-temporal correlations in patient demand, which are crucial for making informed decisions in our research. Additionally, a few studies have explored patient transfers between hospitals, motivated by emergent practices during the pandemic, employing robust optimization (Parker et al. 2020) and queueing-based fluid approximation (Chan et al. 2021). We emphasize that nurse transfer presents its own unique challenges compared to patient or equipment transfer, e.g., nurses need to move back to home location after being transferred instead of being transferred again (in contrast to equipment that can be continuously moved). In addition, we need to design efficient and scalable algorithm to ensure practical implementation rather than treating it solely as a mathematical optimization problem.

Contributions. To the best of our knowledge, this work represents a pioneering implementation that leverages state-of-the-art predictive and prescriptive analytics to optimize nurse staffing at an unprecedented scale. Our focus is on a state-wide program that dynamically reallocates nurses across a hospital network, resulting in substantial contributions to both theory and practice:

- Predictive innovation: We build a novel generative modeling framework that captures the dependence structure and the time-dynamics among census, arrivals, discharges, and underlying latent variables. We design a new temporal-based variational family along with customized encoder-decoder structures for the learning, which provides both efficient representations of the census

timeseries and generates distributional information for the decision-optimization. Comparing to general-purpose prediction methods in machine learning area, we integrate domain knowledge by embedding the patient flow dynamics into the VAE framework. This allows our model to be interpretable and importantly, provides a doubly-stochastic patient census process structure for prescribing optimal decisions in the decision-support phase.

- Prescriptive innovation: We formulate a stochastic optimization (SO) program to effectively capture essential tradeoffs in our nurse deployment program while considering realistic implementation constraints. To efficiently solve the SO, we develop an innovative solution by transforming the original large-scale problem into a tractable linear programming (LP) through a quasi-Monte Carlo method for scenario generation. At the heart of our approach lies a seemingly complex modeling structure: doubly-stochastic processes driven by multivariate Gaussian (MVG) latent variables. This structure not only enhances prediction accuracy but also greatly facilitates the optimization via the feasibility of using a Quasi Monte Carlo method, seamlessly integrating both prediction and optimization components. This integrated design is not only innovative from a methodological perspective, but also critical in providing a scalable solution that can be readily implemented by our partner.
- Implementation: Unlike prior research, which primarily focused on small-scale staffing optimization within individual units or hospitals, our work extends beyond those boundaries. We tackle the complex task of optimizing nurse staffing across an entire state. Implementing decision analytics at such a large scale in the nursing industry, which has been slow to adopt technology, is a significant contribution to this industry and the broader society. Delta Coverage decision analytics offers a transformative solution, enabling nursing organizations to align their staffing needs with accurate and reliable forecasts and results in a substantial reduction in understaffing. Moreover, the implementation of decision analytics in the traditionally technology-resistant industry represents a paradigm shift towards a more data-driven and evidence-based approach to management. This transition can foster a culture of continuous improvement and innovation, unlocking untapped potential and enabling informed decision-making. It also has a far-reaching impact on improving nurse working conditions, enhancing patient outcomes, and especially benefiting underserved regions that are disproportionately affected by nursing shortages.

2. Generative Modeling to Predict Correlated Hospital Occupancies

To overcome challenges associated with existing time-series forecast such as the lack of distributional information and lack of the flexibility to deal with highly time-varying dynamics and nonlinear correlations, we build a generative modeling framework. This framework is based on Li et al. (2023), which developed a novel variational auto-encoding (VAE) method for temporal-based

generative model learning. We tailor and adapt this framework to our specific hospital census prediction setting. The adaptation captures the dependence structure and the time-dynamics among census, arrivals, discharges, and sequence of underlying latent variables. We specify this adapted generative model framework first and then highlight its advantage over existing methods.

2.1. Model Overview

Consider a time-series sequence $\{X_t, t = 0, 1, \ldots, T\}$ with the length of $T+1$, where $X_t \in \mathbb{R}^k$ is a vector that corresponds to the patient census (number of patients) on day t in k hospital units. We denote this sequence as $X_{0:T}$. Our goal is to learn the joint distribution $p(X_{0:T})$. The hospital census is driven by the daily number of arrivals A_t and daily discharges D_t, which are further driven by some underlying "environmental factors" modeled as latent variables. Take the pandemic as an example: the latent variables correspond to the disease spread and recovery, which drive the number of patients that will be hospitalized (arrival) and how long they need to be hospitalized (discharges). To capture this dependence, we adopt the generative modeling framework. Starting with $X_0 = x_0$, the relationship of X_t, X_{t-1}, A_t, D_t can be described recursively as

$$X_t = X_{t-1} + A_t - D_t + \epsilon, \quad t = 1, \ldots, T, \tag{1}$$

where $\epsilon \sim N(0, \tau)$. The sequences of $\{A_t\}$ and $\{D_t\}$ are further driven by the latent sequences $\{Z_t^a\}$ and $\{Z_t^d\}$, respectively. The dependence between the arrival or discharge sequence and the latent sequence can be modeled via some stochastic differential equations (SDE). As we elaborate below, we do not directly learn the arrival or discharges, and thus, we leave the specification of these SDE to Appendix A. Note that the assumption for the normal distribution of X_t's is motivated from the offered-load approximation in queueing networks, which are commonly used to capture the distribution of customer count (census) in service systems (Green et al. 2007).

Cumulative Difference Learning. A common way to learn the joint distribution of $\{X_t\}$ via the generative modeling framework is through step-wise learning, i.e., learning the conditional distribution $X_t | X_{0:t-1}$ recursively for each day t. This method faces an issue: the potential accumulation of errors. That is, for each time step $\ell < T$, if we have a highly inaccurate estimation for the census vector X_ℓ, it will cause the estimations for all the censuses from $\ell + 1$ to time T to deviate significantly from the true values. Because the calculation of census is based on step-wise learning, i.e., X_t depends on X_{t-1}. In other words, the errors accumulate over time, and this could lead to significant deviations from the "truth" for censuses in the distant future.

To overcome this issue, we use a novel cumulative difference learning, specified as follows. First, we use $\Delta_t = A_t - D_t$ to denote the difference between arrival and discharge variables A_t and D_t (i.e., the net changes in X_t's). Then, we define a new variable that captures the cumulative difference:

$$\Gamma_t = X_t - X_0 \; = \; \sum_{i=1}^{t} \Delta_i \; = \; \sum_{i=1}^{t} (A_i - D_i). \tag{2}$$

Here, Γ_t is the cumulative difference between the census on day t and the initial census $X_0 = x_0$. From (1), the relationship between X_0, X_t, and Γ_t can be characterized as:

$$X_t = X_0 + \Gamma_t + \epsilon_t, \quad \epsilon_t \sim N(0, \tau_t), \quad t = 1, \ldots, T. \tag{3}$$

This cumulative difference can be observed by $\gamma_t = x_t - x_0$ (which includes the noise) in the data, where we use lowercase letters to denote the realized/observed values. The noise term ϵ_t capture the measurement errors, which is assumed to follow a multivariate normal distribution with zero mean and covariance τ_t. Note that $X_t \in \mathbb{R}^k$ is a multi-dimensional vector for the census in k locations, hence, the covariance matrix $\tau_t \in \mathbb{R}^{k \times k}$. The covariance matrix is time-varying as the noise ϵ_t for the cumulative difference changes over time.

Following the literature on deep generative models, we assume that the cumulative difference sequence depends on the sequence of latent variables $\{Z_t\}$ through a set of SDE:

$$\begin{aligned} \Gamma_0 &= \Delta_0 = a_0 - d_0, \\ \Gamma_t &= \Gamma_{t-1} + b_t(\Delta_{t-1}) + \sigma_t Z_t, \quad t = 1, \ldots, T. \end{aligned} \tag{4}$$

Here, $Z_0, \ldots, Z_k \sim N(0, I_d)$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^d$, with the unknown parameters to be learned as the drift functions $b_t(\cdot)$, the diffusion matrix σ_t, and the covariance matrix τ_t. The equation (4) can be seen as a discrete-time version of the Cox–Ingersoll–Ross process.

VAE Learning Framework. To learn the unknown parameters for the cumulative difference in (4), we maximize the log-likelihood of joint distribution $p(\gamma_{1:T})$:

$$\log p_\theta(\gamma_{1:T}) = \log \int p_\theta(\gamma_{1:T} \mid z_{1:T}) p(z_{1:T}) dz_{1:T}, \tag{5}$$

where $z_{1:T} = (z_1, \ldots, z_T)$ denote the sequence of realized latent (prior) variables, sampled from the prior distribution $p(z_{1:T}) \sim N(0, I_d)$, $\gamma_{1:T} = \{\gamma_1, \ldots, \gamma_T\}$ is the observed cumulative difference sequence from data, and θ represents parameters in the conditional distribution for $\gamma_{1:T}|z_{1:T}$. The likelihood function is intractable and hard to be evaluated numerically. We adopt the VAE framework for the learning task. At a high level, VAE optimizes the Evidence Lowerbound (ELBO) as the surrogate objective, which contains two major components: (i) learn the conditional distribution $p_\theta(\gamma_t|z_{1:t})$ via a *decoder* $f_\theta(\cdot)$ with parameter θ; (ii) learn $q_\phi(z_{1:T}|\gamma_{1:T})$, which is the variational distribution parameterized with $f_\phi(\cdot)$ with parameter ϕ and approximates the true posterior distribution. Part (i) is called the decoder as it decodes the latent variables $z_{1:t}$ to generate γ_t, while the variational distribution in part (ii) is called the *encoder* as it encodes observed $\gamma_{1:t}$ into the latent space via the variational distribution $q_\phi(z_{1:T}|\gamma_{1:T})$. We design a new temporal-based variational family along with customized encoder-decoder structures for the VAE. The complete details of the ELBO, and the design of the encoder and decoders are delegated to Appendix A. Figure 5 characterizes the entire pipeline for the training and generation procedure.

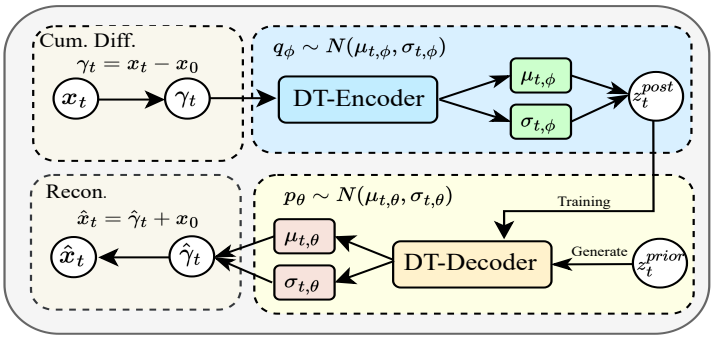
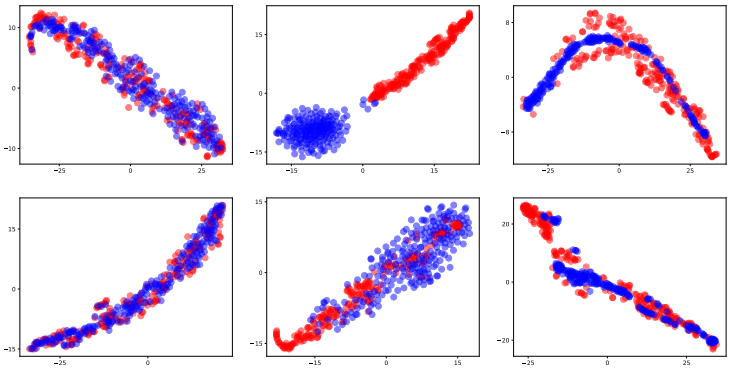


Figure 5 **The overall architecture of DT-VAE with its training and generation procedure. The encoder q_ϕ encodes input data to the latent space, while the decoder p_θ generates data from both encoder samples during training and a prior distribution during generation.**

2.2. Advantages over Existing Tools and Numerical Performance

Our generative-based prediction model offers several advantages over conventional models. First, compared to traditional time-series forecast model such as ARIMA, the encoder-decoder structure provides great flexibility to represent complex functional forms and allows for the easy addition of useful auxiliary covariates to facilitate predictions, e.g., day-of-week or holiday indicators. In particular, this flexible design enables the capture of highly nonlinear and complex spatial-temporal correlations that are difficult to model using conventional statistical methods. This is achieved through the recursive relationship in (4) and the mapping from $Z_{1:t}$ to Γ_t (captured via the decoder f_θ), as Γ_t is correlated with all previous $\Gamma_{1:t-1}$ due to its dependence on the latent variables $Z_{1:t}$, which also drive the correlations among all locations. See Calatayud et al. (2023) for a similar idea to capture the spatial-temporal correlations in crime incidents without explicitly using the latent variables.

Second, by transforming the original census prediction problem into learning the cumulative difference, our method effectively avoids the error accumulation issue associated with recursive prediction that is commonly by time-series generative model, including many state-of-art models such as TimeGAN (Yoon et al. 2019). Since Γ_t represents cumulative differences, it only requires the initial value X_0 for predicting (reconstructing) X_t, in contrast to the recursive reconstruction method used in step-wise learning. In other words, the mapping directly connects $Z_{1:t}$ to all $\Gamma_{1:t}$'s at once. Any bias present in the reconstructed Γ_{t-1} will not impact Γ_t since it is solely determined by the latent variables. See Figure 6 for a comparison with benchmark algorithms, which shows the advantage of our algorithm in addressing these issues. Moreover, we essentially achieve a *conditional independence* among the census conditioning on the latent variables. This will further play a crucial role in facilitating the prescriptive (decision optimization) part, which will be discussed in Section 4.



(a) DT-VAE (b) T-VAE (c) TimeGAN

Figure 6 **t-SNE visualization: Our algorithm (a), Naive time-series VAE (b), and TimeGAN (c) for census generation in a hospital's Med/Surg and ICU units. Red denotes original data, blue denotes generated data. Better mixing of the dots indicates higher quality generated data.**

3. Stochastic Optimization for Network Decision Making

We built a two-stage stochastic optimization (SO) that takes the forecast as input and generates on-call and deployment decisions over a three-week horizon, implemented in a "closed loop" rolling-horizon manner. At the beginning of each week, based on a 21-day forecast, this SO prescribes the weekly schedule on how many nurses to put on call for potential deployment 1-2 weeks in advance (Step 1 of the DSS). Then, at the beginning of each day, based on the realized census and updated forecast for the rest of the week, we re-solve the SO and use the first-day decision to determine the actual deployments on the current day (Step 2 of the DSS).

The primary objective is to reduce system-wide understaffing without being too disruptive to nurse's lives through excessive or unreasonable travel schedules. We denote the on-call decision as $\mathbf{a} = \{a_t^{ij}\}$, where each a_t^{ij} is the number of DC nurses to put on-call for a future transfer from unit i to unit j on day t. Similarly, we denote the recourse call-in decision as $\mathbf{b} = \{b_t^{ij}\}$, which is made after seeing the realization of the census sample path $\mathbf{X} = \{X_t^i\}$. The recourse decision corresponds to either activates the transfer of an on-call nurse or cancels the transfer. The transferred nurse is committed to work on multiple shifts at unit j for a length of S^{ij} days, referred to as the *secondment*. The secondment is an important design feature that avoids nurses having to travel two long-distance legs in addition to working a 12-hour shift and ensures continuity of care.

To capture multiple tradeoffs that are used as program design parameters determining features like the efficacy for the system and attractiveness to DC nurses, we consider the following costs. Nurse shortage is captured via the understaffing cost. The cost associated with the transfer decision $\mathbf{a}$ comes from two parts: (i) the fixed cost that compensates for the transfer, c_t^{ij}, which depends on the transfer distance, and (ii) the variable cost that compensates for the length of the secondment $c_p S^{ij}$. If a transfer is cancelled during recourse, we recoup $1-\eta$ percent of the transfer cost. We

set the unit understaffing cost to be 1, and normalize other costs. These costs are not necessarily the actual financial costs (e.g., of premium pay), but rather tuning parameters. During program design we adjusted these costs to achieve desired system performance (see more in Section 5).

Mathematically, the objective is

$$\min_{\mathbf{a}} \quad \sum_{t=1}^{T}\sum_{i=1}^{k}\sum_{j=1}^{k}(c_t^{ij}+c_pS^{ij})a_t^{ij}+\mathrm{E}_{\mathbf{X}}\left[V\left(\mathbf{a},\mathbf{b},\mathbf{X}\right)\right], \tag{6}$$

$$V\left(\mathbf{a},\mathbf{b},\mathbf{X}\right)=\min_{\mathbf{b}}\sum_{t=1}^{T}\sum_{i=1}^{k}\Big[\left(X_t^i-\bar{n}_t^i\right)^+-(1-\eta)(c_t^{ij}+c_pS^{ij})\left(a_t^{ij}-b_t^{ij}\right)^+\Big], \tag{7}$$

subjected to

$$\sum_{j=1}^{k}a_t^{ij}\le d_t^i-\sum_{j=1}^{k}\sum_{\ell=(t-S^{ij}+1,1)^+}^{t-1}a_\ell^{ij}, \quad \sum_{j=1}^{k}b_t^{ij}\le\sum_{j=1}^{k}a_t^{ij}, \quad \forall i,t, \tag{8}$$

where d_t^i is the number of available DC nurses with home location i on day t, and $\bar{n}_t^i$ is the number of nurses available at location i on day t after considering the actual deployment (recourse decision) and secondment to the number of scheduled regular nurses n_t^i:

$$\bar{n}_t^i=n_t^i-\sum_{j=1}^{k}\sum_{\ell=t-S^{ij}}^{t}b_\ell^{ij}+\sum_{j=1}^{k}\sum_{\ell=t-S^{ji}}^{t}b_\ell^{ji}. \tag{9}$$

Note that the nurse demand and the patient census are not equal, as the former is adjusted based on the patient-nurse ratio, e.g., one nurse is required for taking care of two patients in ICU or four patients in Med/Surg units. For notational simplicity, we use the same $\mathbf{X}$ to denote the nurse demand and the patient census throughout the paper. For brevity, we do not specify the full set of constraints that are employed in the implementation, particularly during the actual deployment stage. During the iterative design process, we utilized additional constraints and parameter tuning to capture design specifications, such as: (1) limiting the fraction of time that a nurse is put on-call but not called in, (2) limiting the average daily volume of nurses working remote shifts, (3) ensuring that nurses do not take two travel assignments in a row without working in their home hospital in between, and (4) ensuring equitable use of Delta Coverage deployments to avoid perceived (or real) favoritism for certain hospitals.

4. Integration of Predictive and Prescriptive Components

The most difficult part in the objective function (6) is the expectation $\mathrm{E}_{\mathbf{X}}\left[V\left(\mathbf{a},\mathbf{b},\mathbf{X}\right)\right]$. To evaluate this expectation, a common approach is to use sample-average method. In our setting, the sampling-based optimization should fully account for the generative modeling structure used in the forecast part as specified in Section 2. That is, instead of directly sampling the census sequence X's as in

conventional settings, we first sample the latent sequence Z's from multivariate standard Gaussian distribution. Then, conditional on each sampled latent sequence $z = z_{1:T}$, we obtain the mean and covariance for $X|z$ via the decoder and sample accordingly.

Recall that a benefit of our generative predictive framework is that we achieve a conditional decomposition both temporally and spatially since the spatial-temporal correlations are captured via the decoder f_θ. Specifically, conditional on a sampled (realized) sequence $z_{1:T}$, the mean for the census in unit i on day t is $\mu^i_{t,\theta}$, and the variance is $\sigma^i_{t,\theta}$. For a given initial census $X_0 = x_0$, each X^i_t can be characterized as

$$X^i_t \sim \big(x_0 + \mu^i_{t,\theta}(z_{1:t})\big) + \sigma^i_{t,\theta}(z_{1:t}) \cdot N(0,1), \quad t = 1,\dots,T, i = 1,\dots,k,$$

where $N(0,1)$ is a standard normal r.v.. In other words, X^i_t is a doubly-stochastic r.v. that depends on the latent variables $z_{1:t}$ and $\zeta_{i,t} \sim N(0,1)$. For the doubly-stochastic r.v., sample-average methods require two loops to obtain the samples, where the outer loop is to sample the latent variables and the inner loop is to sample the normal r.v. $\zeta_{i,t}$'s. In the following, we let $\zeta = \{\zeta_{i,t}\}$ for the set of i.i.d normal r.v.'s for each station i and each day t, used in conjunction with $z_{1:t}$ to create the doubly-stochastic distribution of X^i_t. Let $z^m_{1:t}$ be the m^{th} sample of the latent sequence, and ζ^ℓ be the ℓth set of sampled r.v.'s.

In the interest of space, we focus on writing the calculation of the under-staffing part in $\mathrm{E}_{\mathbf{X}}\left[V(\mathbf{a},\mathbf{b},\mathbf{X})\right]$. We define $y^{m,\ell}_{i,t}$ as the auxiliary variable that approximates the value of the under-staffing function in unit i on day t given the mth sample $z^m_{1:t}$ and the ℓth sample ζ^ℓ:

$$\mathrm{E}_{\mathbf{X}}\left[\sum_{t=1}^{T}\sum_{i=1}^{k}(X^i_t - \bar{n}^i_t)^+\right] \approx \frac{1}{M\cdot L}\sum_{m=1}^{M}\sum_{\ell=1}^{L}\sum_{t=1}^{T}\sum_{i=1}^{k} y^{m,\ell}_{i,t}, \tag{10}$$

$$s.t.$$

$$y^{m,\ell}_{i,t} \geq \big(x_0 + \mu^i_{t,\theta}(z^m_{1:t})\big) + \sigma^i_{t,\theta}(z^m_{1:t})\cdot\zeta^\ell_{i,t} - \bar{n}^i_t, \qquad \forall i,t,\ell,m, \tag{11}$$

$$y^{m,\ell}_{i,t} \geq 0, \qquad \forall i,t,\ell,m. \tag{12}$$

Here, for ease of exposition, we suppress the dependence of $\bar{n}^i_t$ on the recourse decision, which can reduce the understaffing through the minimization in $V(\mathbf{a},\mathbf{b},\mathbf{X})$; see (6) and (9).

Efficient Sampling. For both the inner and outer loops, we need to sample from a multivariate standard Gaussian distribution (for $z_{1:T}$ and ζ, respectively) to evaluate the sample average in (10). The benefit is that there is no correlation among these Gaussian r.v.'s (as opposed to directly sampling from $\{X^i_t\}$'s), thus, we can sample each coordinate independently. The disadvantage is that the dimension is still high (e.g., $z_{1:T}$ has 21 dimensions when we plan for three weeks out with $T = 21$). Conventional Monte Carlo method is a viable approach for high-dimensional space but

suffers from larger variance, requiring a large number of samples to achieve accurate evaluation of the sample average. This imposes a great computational challenge for our healthcare partners as the open-source optimization solver cannot handle a large number of samples. To address this issue, we leverage the Quasi Monte Carlo method (QMC), which is known to reduce variance in sampling: it can improve the rate of convergence from $O(1/\sqrt{M})$ in conventional MC method to $O(1/M)$ in QMC, where M is the number of samples (Caflisch 1998). This means a much smaller number of samples is required to achieve similar level of accuracy.

Specifically, we use a variant of the Latin Hypercube Sampling (LHS) (Owen 1998). For a desired number of M samples, we first divide the real line for each coordinate (a univariate Gaussian) into a few adjacent intervals defined via $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$, i.e., a set of M disjoint partitions of $\mathbb{R}$. For $m = 1, \ldots, M$, $\int_{\mathcal{I}_m} \phi(x)dx$ is the integral of the density in each partition $\mathcal{I}_m$ with $\phi(x)$ being the pdf of the standard Gaussian. We choose the partition such that each $\int_{\mathcal{I}_m} \phi(x)dx = 1/M$ is equal, and we set a "representative value" u_m for partition m using the middle point of $\mathcal{I}_m$. Finally, we follow the LHS method to create M samples, i.e., we create T independent and random permutations of the vector $u = \{u_1, \ldots, u_M\}$ and match the value from each of the T coordinates to have M sampled vectors of T dimensions, $\{z^m_{1:T}\}_{m=1}^M$. We create the samples $\{\zeta^\ell\}_{\ell=1}^L$ in a similar way.

Notably, even though our method still requires sampling from a high-dimensional space, the QMC method allows us to sample efficiently regardless of the dimensions, reducing sampling variance and the number of samples needed. This is equivalent to adding *carefully chosen* cuts to the LP, as opposed to relying on purely random-generated cuts from the MC method (traditional sample average method), to achieve more accurate approximation in (10) and speedup the solution. The feasibility of using the LHS method benefits greatly from the multivariate Gaussian distribution, as it allows for an explicit form of the pdf and independent sampling for each dimension. This advantage would not be possible if working directly with the census variable, given the complexity of the joint pdf and the correlations. In addition, the sample is from the multivariate standard Gaussian (instead of the census variable), which can be *re-used* to avoid re-sampling from X when the forecast is updated and optimization is re-solved each day (Step 2 in the DSS). The mapping from Z to X is an exogenous input that can be trained offline (e.g., on better computational platform) and loaded as a matrix to the LP with warm start techniques to significantly increase solution speeds.

In summary, we transform a large-scale SO problem into a tractable LP using QMC. The seemingly complex generative framework actually enhances both prediction and prescription capabilities. This integration highlights the significance of the generative framework, while also providing a portable solution for our partner's real-world implementation needs.

5. Implementation of Delta Coverage and Practical Challenges

In this section, we outline our tiered implementation process and discuss the challenges encountered when deploying an analytics-based solution in a healthcare environment, which may also apply to other researchers in similar endeavors.

5.1. Implementation

Our co-developed Delta Coverage Analytics Suite, launched in Oct 2021 as a Microsoft Power BI application, underwent three phases of implementation. In the first phase of implementation, we logged the performance of the recommendations from October 2021 to March 2022, with the tool running each day in real time to provide proof of value as well as strategic insights for program adoption. Our analysis revealed system-wide improvements across all metrics, including a 5% reduction in understaffing and a 1% reduction in overstaffing. In the second phase, we collaborated closely with the management team to iteratively design and enhance feasible deployment regions. This involved crafting appropriate incentive mechanisms to encourage nurse participation and using the model to optimize program logistics in response to valuable feedback from nursing leadership at individual hospitals and prospective nurses; see more details of the implementation process in Section 5. In the third phase, the pilot was approved and officially launched in May 2023 with 10 Delta Coverage nurses and 6 participating hospitals.

Performance Log. We built a system that automatically logs all data pulled for input into the optimization and forecast models as well as the outputs of those models. This log is updated each time the DC dashboard is run since some of the data cannot be collected after the fact; e.g., data that comes from central data warehouses can get overwritten with newer data. This log has allowed us to detect changes in the enterprise data systems that could affect our model inputs, validate forecast accuracy, and monitor the value of the program to the nursing organization.

Tiered Implementation. Due to the novelty of the program we had no benchmark examples of implementing such a program. To mitigate potential risks, we executed a tiered implementation with report-outs to gain buy-in from upper-level management after each phase.

Pre-implementation: Historical Counterfactual. Before implementation, we conducted a counterfactual analysis using 2 months of historical data and estimated a 4% reduction in understaffing by implementing the optimal recommendations (we did not measure overstaffing). This "low-cost" testing of the analytics suite was crucial in gaining management buy-in, as it demystified the "blackbox" DSS and showcased the power of OR analytics. This was especially valuable given previous experiences with consulting companies that provided opaque solutions lacking actionable information.

Phase 1: Live-test Run. Based on the promising results, we launched Phase 1, building a PowerBI Dashboard and integrating it with IUH data warehouses and the analytics suite. Over the next 5

months, we field-tested the system live, running it daily to estimate the FTEs needed for support and maintenance. The results showed a 5% reduction in understaffing and a 1% reduction in overstaffing. These outcomes, along with strong advocacy from nursing organization leadership, convinced broader executive leadership to support a pilot.

Phase 2: Iterative Design Improvement. A critical factor in the success of our iterative design process was the ability to use our stochastic optimization and census forecast model to instantly project the impact of different design decisions. The optimization also has tuning parameters that can ensure the program is operationalized to meet target specifications.

Phase 3: Pilot Program. We began by identifying a group of hospitals to participate in the pilot through discussions with all of IU Health's Chief Nursing Officers (CNOs). Subsequently, we sought feedback from the CNOs of the participating hospitals and iterated multiple times to design a program that would be conducive to adoption. The recruitment process for the DC nurse pool was a crucial aspect of the pilot program, requiring considerable effort to attract highly skilled and location-flexible nurses. These nurses not only needed to be willing to travel but also had to be able to work in multiple clinical settings, transcending single specialties, acuity levels, or units. Several program specification redesigns were necessary to achieve the recruitment target, and by the program's launch on May 1, 2023, we successfully recruited 10 DC nurses both internally and externally to IUH. The reasons behind the delay between the prototype and launch, along with other practical challenges, are described in the next section.

5.2. Practical Challenges and Lessons Learned

Nursing Crisis. The greatest challenge to the Delta Coverage program was, ironically, the primary impetus for the program itself: the nursing shortage crisis. By October 2021, we had a fully-functional prototype of the DC Dashboard which we completed testing in April 2022. However, the pilot launch was delayed until May 2023 due to the unprecedented severity and duration of the nursing shortage crisis in Indiana. During this period, the National Guard had to be called in multiple times to support hospital staffing across the state.

While the delay in the pilot launch seemed ironic, it is crucial to recognize that the crisis highlighted the urgent need for innovative solutions like the Delta Coverage program. The gap between the prototype development and the pilot launch provided the opportunity for the academic team to refine and strengthen the supporting analytics theory. Additionally, the DC analytics suite proved its value during the crisis, providing critical insights and support to IUH in managing the nursing shortage at their hospitals. This demonstrated the suite's versatility and effectiveness, even in addressing challenges beyond the DC program's original scope.

Despite the challenges posed by the nursing shortage crisis, the collaboration between the academic team and IUH remained strong. The continuous communication and development efforts

allowed us to further enhance the DC program's capabilities and ensure its readiness for the pilot. The experience gained during the crisis response has enriched our understanding of the healthcare environment and reaffirmed the value and potential impact of the Delta Coverage program in effectively managing nurse shortages in the future. In January 2023, the team decided to restart planning for the pilot launch, focusing on two major milestones: relaunching and retesting the analytics suite, and recruiting nurses for the Delta Coverage program.

DC Analytics Suite. When we began the relaunch, we encountered several changes in the underlying data systems, including modifications to enterprise data systems that impacted our data pipeline, acuity reclassification in different units, and the second-largest hospital at IUH that had not yet been reintegrated into the central data warehouse after relocating to a new building. Despite identifying and addressing these issues, the forecast and optimization continued to perform well after a year of dormancy. Another significant data challenge we faced, common to many hospitals developing data-driven operational analytics, was that hospital data is primarily designed for billing and finance. This required us to implement major workarounds to ensure accurate operational conclusions. For example, we had to use patient location data (the location where the patient is billed) to construct hospital occupancy data. However, we discovered a double counting issue with numerous patients, where they were mistakenly counted in two places due to the inpatient bed being held for the patient while they were in surgery or a recovery room. Our team addressed these challenges through advanced planning, anticipating future changes (e.g., hospital moves), and incorporating an automated change detection mechanism.

Recruitment. As mentioned, one of the major challenges and milestones was recruiting nurses for this novel job description. This involved both ingenuity and due diligence from the nursing organization management as well as scenario testing and operational design using the analytics engine. Despite the well-planned and well-executed iterative design process, we were unable to recruit a sufficient number of qualified nurses on our first attempt. In the subsequent redesign we were able to use the tunable model hyper-parameters to include additional desirable features mentioned by the different nursing teams through a second iterative process. This involved identifying different design specifications that would make the program more attractive to DC nurses and features that ensured fairness among hospitals and among DC nurses. Another feedback mechanism was information sessions for DC eligible nurses. Other design changes tested in the analytics suite include partitioning the network into smaller travel zones (or pods) each with its own set of DC nurses, enforcing limits on the probabilities that a nurse would be deployed from the on-call list, adjusting length of travel secondments (how many shifts a Delta Coverage nurse works at a remote location), limiting the fraction of shifts that a DC nurse works at a remote hospital, and ensuring the fraction of DC shifts allocated to each participating hospital was fair among others. The second wave of

recruitment proved to be a success, thanks to the implementation of design changes tested in the analytics suite.

6. Conclusion

The state-wide Delta Coverage Program, a collaborative effort between academia and industry, represents a groundbreaking solution for addressing nurse staffing challenges. With its integrated predictive-prescriptive framework, the Delta Coverage Analytics Suite provides real-time distributional nurse demand forecasts and dynamic deployment decisions, resulting in reduced understaffing, optimized resource utilization, and improved nurse job satisfaction and patient care quality. The successful pilot phase showcased significant reductions in understaffing and overstaffing, demonstrating its potential for long-term impact in mitigating nurse shortages and burnout, especially in underserved regions. This pioneering program offers a sustainable solution to address the multifaceted challenges of nurse staffing, burnout, and healthcare disparities, fostering a nurturing environment for nurses and strategically allocating resources. The program's positive impact extends beyond immediate staffing concerns, leaving a lasting impression on the well-being of the nursing workforce and the communities they serve.

References

Aiken LH, Sloane DM, Bruyneel L, Van den Heede K, Griffiths P, Busse R, Diomidous M, Kinnunen J, Kózka M, Lesaffre E, et al. (2014) Nurse staffing and education and hospital mortality in nine european countries: a retrospective observational study. *The lancet* 383(9931):1824–1830.

Allen LJ (2017) A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling* 2(2):128–142, ISSN 2468-0427, URL `http://dx.doi.org/https://doi.org/10.1016/j.idm.2017.03.001`.

Allen LJS (2008) *An Introduction to Stochastic Epidemic Models*, 81–130 (Berlin, Heidelberg: Springer Berlin Heidelberg), ISBN 978-3-540-78911-6, URL `http://dx.doi.org/10.1007/978-3-540-78911-6_3`.

Anderson D, Bjarnadottir MV, Nenova Z (2022) Machine learning in healthcare: Operational and financial impact. *Innovative Technology at the Interface of Finance and Operations: Volume I* 153–174.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.

Blegen MA, Goode CJ, Spetz J, Vaughn T, Park SH (2011) Nurse staffing effects on patient outcomes: safety-net and non-safety-net hospitals. *Medical care* 406–414.

Caflisch RE (1998) Monte carlo and quasi-monte carlo methods. *Acta numerica* 7:1–49.

Calatayud J, Jornet M, Mateu J (2023) Spatio-temporal stochastic differential equations for crime incidence modeling. *Stochastic Environmental Research and Risk Assessment* 1–16.

Chan T, Pogacar F, Sarhangian V, Hellsten E, Razak F, Verma A (2021) Optimizing inter-hospital patient transfer decisions during a pandemic: A queueing network approach. *Available at SSRN 3975839* .

Cox JC, Ingersoll Jr JE, Ross SA (2005) A theory of the term structure of interest rates. *Theory of valuation*, 129–164 (World Scientific).

Desai A, Freeman C, Wang Z, Beaver I (2021) Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095* .

Esteban C, Hyland SL, Rätsch G (2017) Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* .

Flinkman M, Leino-Kilpi H, Salanterä S (2010) Nurses' intention to leave the profession: integrative review. *Journal of advanced nursing* 66(7):1422–1434.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* 16(1):13–39.

Griffiths P, Saville C, Ball J, Jones J, Pattison N, Monks T, Group SNCS, et al. (2020) Nursing workload, nurse staffing methodologies and tools: A systematic scoping review and discussion. *International Journal of Nursing Studies* 103:103487.

Hu Y, Chan CW, Dong J (2021) Prediction-driven surge planning with application in the emergency department. *Submitted to Management Science* .

Li T, Wu C, Shi P, Wang X (2023) Cumulative difference learning vae for time-series with temporally correlated inflow-outflow. Working Paper.

Mogren O (2016) C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904* .

Owen AB (1998) Latin supercube sampling for very high-dimensional simulations. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 8(1):71–102.

Parker F, Sawczuk H, Ganjkhanloo F, Ahmadi F, Ghobadi K (2020) Optimal resource and demand redistribution for healthcare systems under stress from covid-19. *arXiv preprint arXiv:2011.03528* .

Saville CE, Griffiths P, Ball JE, Monks T (2019) How many nurses do we need? a review and discussion of operational research techniques applied to nurse staffing. *International journal of nursing studies* 97:7–13.

Shi P, Helm JE, Chen C, Lim J, Parker RP, Tinsley T, Cecil J (2023) Operations (management) warp speed: Rapid deployment of hospital-focused predictive/prescriptive analytics for the covid-19 pandemic. *Production and Operations Management* 32(5):1433–1452.

Spetz J (2021) Leveraging big data to guide better nurse staffing strategies.

Yoon J, Jarrett D, Van der Schaar M (2019) Time-series generative adversarial networks. *Advances in neural information processing systems* 32.

Zlotnik A, Gallardo-Antolin A, Alfaro MC, Pérez MCP, Martínez JMM, et al. (2015) Emergency department visit forecasting and dynamic nursing staff allocation using machine learning techniques with readily available open-source software. *CIN: Computers, Informatics, Nursing* 33(8):368–377.

Appendix A: More Details on Generative Model

A.1. SDE for Modeling Generative Dependence Structure

Motivated by the stochastic SIR model Allen (2008, 2017), we assume that the arrivals A_t and discharges D_t follow

$$\begin{aligned} A_0 &= a_0; \quad D_0 = d_0; \\ A_t &= A_{t-1} + b_a(A_{t-1}) + \sigma_a Z_t^a, t = 1, \ldots, T \qquad (13) \\ D_t &= D_{t-1} + b_d(D_{t-1}) + \sigma_d Z_t^d, t = 1, \ldots, T \qquad (14) \end{aligned}$$

where the sequences of latent variables $Z_1^a, \ldots, Z_T^a \sim^{iid} \mathcal{N}(0, I_k)$ and $Z_1^d, \ldots, Z_T^d \sim^{iid} \mathcal{N}(0, I_k)$ are all i.i.d. standard Gaussian vectors in $\mathbb{R}^k$ and drive the arrival and discharge processes. Equations (13) and (14) can be seen as the discretized version of the original stochastic differential equations for the stochastic SIR model, with $b_a(\cdot)$ and $b_d(\cdot)$ as the (unknown) drift functions and $\sigma_a Z_t^a$ and $\sigma_d Z_t^d$ as the (unknown) diffusion terms.

A.2. VAE Learning Framework

Instead of directly evaluating the likelihood function $p_\theta(\gamma_{1:T})$ given in (5), VAE optimizes the Evidence Lowerbound (ELBO) as the surrogate objective, derived as below in our setting:

$$\begin{aligned} \log p_\theta(\gamma_{1:T}) &= \log \int p_\theta(\gamma_{1:T}, z_{1:T}) dz_{1:T} \\ &= \log \int p_\theta(\gamma_{1:T}, z_{1:T}) \frac{q_\phi(z_{1:T} \mid \gamma_{1:T})}{q_\phi(z_{1:T} \mid \gamma_{1:T})} dz_{1:T} \\ &\geq \mathbb{E}_{z_{1:T} \sim q_\phi} \left[\log \left(\frac{p_\theta(\gamma_{1:T}, z_{1:T})}{q_\phi(z_{1:T} | \gamma_{1:T})} \right) \right] \\ &= \mathbb{E}_{z_{1:T} \sim q_\phi} \left[\log \left(\frac{\prod_{t=1}^T p(\gamma_t | z_{1:t})) p(z_t | z_{1:t-1})}{\prod_{t=1}^T q_\phi(z_t | z_{1:t-1}, \gamma_{1:t})} \right) \right] \qquad (15) \\ &= \sum_{t=1}^T \mathbb{E}_{z_{1:t}} \log p(\gamma_t | z_{1:t}) \\ &\quad - \mathbb{E}_{z_{1:t-1}} D_{KL} \Big(q_\phi(z_t | z_{1:t-1}, \gamma_{1:t}) || N(0, I) \Big) \\ &= \mathcal{L}(\gamma_{1:T}). \end{aligned}$$

Recall that the key for VAE evaluation lies in two parts. The first part is to learn the conditional distribution $p_\theta(\gamma_t | z_{1:t})$ via a *decoder* $f_\theta(\cdot)$ with parameter θ. It is called the decoder as it decodes the latent variables $z_{1:t}$ to generate γ_t. The second part is to learn $q_\phi(z_{1:T} | \gamma_{1:T})$, which is the variational distribution with parameter ϕ that approximates the true posterior distribution. This variational distribution is called the *encoder*, parameterized with $f_\phi(\cdot)$ with parameter ϕ. It encodes observed $\gamma_{1:t}$ into the latent space via the variational distribution $q_\phi(z_{1:T} | \gamma_{1:T})$. In implementation, we use an additional hyperparameter $\lambda > 0$ in front of the KL term to further balance the two parts in ELBO.

In the rest of this section, we will use $f_\theta(z_{1:t})$ and $p_\theta(\gamma_t | z_{1:t})$ interchangeably and use z^{prior} to denote samples from the prior distribution; we will use $f_\phi(\gamma_{1:t})$ and $p_\phi(z_{1:t} | \gamma_{1:t})$ interchangeably, and z^{post} to denote samples from the posterior distribution.

Decoder. A key step in deriving the ELBO in (15), particularly from line 3 to line 4, is via the following decomposition

$$\begin{aligned} p_\theta(\gamma_{1:T}, z_{1:T}) &= p_\theta(\gamma_{1:T}|z_{1:T})p(z_{1:T}) \\ &= \Big(\prod_{t=1}^{T} p_\theta(\gamma_t|z_{1:t})\Big)p(z_{1:T}) = \prod_{t=1}^{T} p_\theta(\gamma_t|z_{1:t}) \prod_{t=1}^{T} p(z_t|z_{1:t-1}), \end{aligned} \tag{16}$$

where $p(z_t|z_{1:t-1})$ denotes the conditional prior distribution for latent variables z_t. We make an important assumption here for the conditional distribution $p_\theta(\gamma_{1:T}|z_{1:T})$ and prior distribution $p(z_{1:T})$. As discussed, in the cumulative difference learning setup, each γ_t depends on latent variables $z_{1:t}$ to avoid error accumulation. This essentially makes γ_t to be conditionally independent across different time steps given realized latent variables $z_{1:t}$. That is, for any two time steps $w \neq v \leq T$, the cumulative difference variable $(\gamma_w|z_{1:w}) \perp (\gamma_v|z_{1:v})$ are independent conditional on corresponding latent variables. This assumption is crucial, allowing the transformation from $p_\theta(\gamma_{1:T}|z_{1:T})$ to the product form $\prod_{t=1}^{T} p_\theta(\gamma_t|z_{1:t})$.

Following the VAE literature, we assume the conditional distribution $p_\theta(\gamma_t|z_{1:t}) \sim N(\mu_{t,\theta}, \sigma_{t,\theta})$, i.e., a multivariate Gaussian distribution with mean $\mu_{t,\theta}$ and diagonal covariance matrix $\sigma_{t,\theta}$ for time t. This is a reasonable assumption in our setting as the *difference* in census can be both positive or negative (in contrast to that arrivals or departures have to be positive). Under the Gaussian assumption, the decoder f_θ is represented by the mean and covariance matrix, denoted as $f_\theta = \{(\mu_{t,\theta}, \sigma_{t,\theta})\}_t$, with the subscript t highlighting the time-dependency. For the prior distribution, we assume they are independent Gaussian, namely, $p(z_t|z_{1:t-1}) \sim N(0, I)$ with $I \in \mathbb{R}^{d\times d}$ being the identity matrix. Though the priors are assumed to be independent, the decoder f_θ allows us to capture the underlying complex correlations.

Encoder. We factor the variational distribution $q_\phi(z_{1:T}|\Gamma_{1:T})$ as

$$q_\phi(z_{1:T}|\gamma_{1:T}) = \prod_{t=1}^{T} q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}). \tag{17}$$

During the training stage, we will sample z_t^{post} from the posterior distribution $q_\phi(z_t|z_{1:t-1}, \gamma_{1:t})$ and let the decoder reconstruct the observed γ_t's. The sampling is recursive as we need to conditional on sampled variables $z_{1:t-1}^{post}$ and observed $\gamma_{1:t}$ when sampling for time t. Following the VAE literature, we assume that variational distribution $q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}) \sim N(\mu_{t,\phi}, \sigma_{t,\phi})$, i.e., a multivariate Gaussian distribution with mean $\mu_{t,\phi}$ and diagonal covariance matrix $\sigma_{t,\phi}$. Under this Gaussian assumption, the encoder f_ϕ is represented by the mean and covariance matrix, denoted as $f_\phi = \{(\mu_{t,\phi}, \sigma_{t,\phi})\}_t$, with the subscript t highlighting the time-dependency.

A.3. Encoder and Decoder Design

In this section, we specify the encoder and decoder designs via recurrent neural networks.

Decoder design. For the generative process, DT-VAE uses a decoder $f_\theta(\cdot)$ with parameter θ to decode latent variables $z_{1:t}$ to generate γ_t. In other words, the decoder $f_\theta(\cdot)$ learns the conditional distribution

$p_\theta(\gamma_t|z_{1:t})$. A key step in deriving the ELBO in (15), particularly from line 3 to line 4, is via the following decomposition for $p_\theta(\gamma_{1:T}, z_{1:T})$:

$$\begin{aligned} p_\theta(\gamma_{1:T}, z_{1:T}) &= p_\theta(\gamma_{1:T}|z_{1:T})p(z_{1:T}) \\ &= \Big(\prod_{t=1}^{T} p_\theta(\gamma_t|z_{1:t})\Big)p(z_{1:T}) \\ &= \prod_{t=1}^{T} p_\theta(\gamma_t|z_{1:t}) \prod_{t=1}^{T} p(z_t|z_{1:t-1}), \end{aligned} \tag{18}$$

where $p_\theta(\gamma_t|z_{1:t})$ denotes the approximation of the true conditional distribution $p(\gamma_t|z_{1:t})$ and $p(z_t|z_{1:t-1})$ denotes the conditional prior distribution for latent variables z_t.

From (18), we make an important assumption on the conditional distribution $p_\theta(\gamma_{1:T}|z_{1:T})$ and prior distribution $p(z_{1:T})$. As previously mentioned, for each γ_t, it solely depends on latent variables $z_{1:t}$ to avoid error accumulation. This essentially makes γ_t to be *conditionally independent* across different time steps given the latent variables $z_{1:t}^{prior}$. That is, for any two time steps $w \neq v \leq T$, the cumulative difference variable $(\gamma_w|z_{1:w}) \perp (\gamma_v|z_{1:v})$ are independent conditional on corresponding latent variables. This assumption is crucial, allowing the transformation from $p_\theta(\gamma_{1:T}|z_{1:T})$ to the product form $\prod_{t=1}^{T} p_\theta(\gamma_t|z_{1:t})$.

Following the VAE literature, we assume the conditional distribution $p_\theta(\gamma_t|z_{1:t}) \sim N(\mu_{t,\theta}, \sigma_{t,\theta})$, i.e., a Gaussian distribution with a diagonal covariance matrix. This also makes sense in our application as the difference in census can be both positive or negative (in contrast to that arrivals or departures have to be positive). Also note that $\sigma_{t,\theta}$ is time-varying as from (3). For the prior distribution, we assume they are independent Gaussian, namely, $p(z_t|z_{1:t}) \sim N(0, I)$. Though z_t^{prior}'s are independent, the decoder f_θ still allows us to capture the underlying correlation via the relationship between γ_t and $z_{1:t}^{prior}$. Specifically, we design the decoder via a recurrent network f_{θ_1}, enclosing all time steps information $z_{1:t}^{prior}$ recursively, with a feedforward network f_{θ_2}, further transforming the input to $\mu_{t,\theta}$ and $\sigma_{t,\theta}$, i.e.,

$$h_{t,\theta_1} = f_{\theta_1}(h_{t-1,\theta_1}, z_t) \qquad (\mu_{t,\theta}, \sigma_{t,\theta}) = f_{\theta_2}(h_t) \tag{19}$$

where h_{t,θ_1} is the hidden state in the RNN structure f_{θ_1}.

Encoder design. Next, we will describe the posterior distribution, also known as the encoder. DT-VAE learns an encoder $f_\phi(\cdot)$ with parameter ϕ to encode observed $\gamma_{1:t}$ into the variational (posterior) distribution $q_\phi(z_{1:T}|\gamma_{1:T})$. We factor the posterior distribution $q_\phi(z_{1:T}|\Gamma_{1:T})$ as

$$q_\phi(z_{1:T}|\gamma_{1:T}) = \prod_{t=1}^{T} q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}) \tag{20}$$

During the training stage, we will sample z_t^{post} from the posterior distribution $q_\phi(z_t|z_{1:t-1}, \gamma_{1:t})$ and let the decoder reconstruct the observed γ_t's. For samples from posterior distribution, at each t, we sample z_t^{post} from the distribution conditioned on the historical posterior variables $z_{1:t-1}^{post}$ and all observed $\gamma_{1:t}$.

Following the VAE literature, we assume that variational distribution $q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}) \sim N(\mu_{t,\phi}, \sigma_{t,\phi})$, i.e., a Gaussian distribution with a diagonal covariance matrix, where the $\mu_{t,\phi}$ and $\sigma_{t,\phi}$ are learned using the encoder f_ϕ. To capture that the historical information relies on both z^{post}'s and γ's, we decompose f_ϕ into three functions with parameters ϕ_1, ϕ_2 and ϕ_3:

$$\begin{aligned} h_{t,\phi_1} &= f_{\phi_1}(h_{t-1,\phi_1}, \gamma_t) \\ \mu_{t,\phi} &= f_{\phi_2}(h_{t,\phi_1}, \mu_{t-1,\phi}) \\ \sigma_{t,\phi} &= f_{\phi_3}(h_{t,\phi_1}, \sigma_{t-1,\phi}) \end{aligned} \tag{21}$$

where h_{t,ϕ_1} is the hidden state in RNN structure f_{ϕ_1}. For each time step, h_{t,ϕ_1} encodes the all observed $\gamma_{1:t}$. The RNN structure f_{ϕ_2} will output the mean of posterior distribution $\mu_{t,\phi}$ by utilizing the h_{t,ϕ_1} and previous $\mu_{t-1,\phi}$. Therefore, for each time step, the current mean $\mu_{t,\phi}$ contains information of previous means $\mu_{1:t-1,\phi}$, which resembles the conditional structure in $q_\phi(z_t|z_{1:t-1},\gamma_{1:t})$ from (20). Similarly, the RNN structure f_{ϕ_3} outputs $\sigma_{t,q}$ by utilizing h_{t,ϕ_1} and $\sigma_{t-1,\phi}$ that contain prior information of $\gamma_{1:t}$ and $z^{post}_{1:t-1}$. It is noteworthy that the recursive design is guided by our mathematical results, which turn out to be critical. We tried other heuristic designs without properly using the prior information as suggested by the theoretical form and they failed to learn, which highlights the importance of theoretical justification.

Appendix B: Prediction Performance Evaluation

We demonstrate the advantage of our method (the generative modeling structure and cumulative difference learning) over traditional statistical methods such as AR model. Our evaluation platform is a semi-synthetic hospital census dataset created from a simulation model, which is calibrated with real data from a partner hospital. Specifically, the daily arrivals $a(t)$ follow the discretized Cox–Ingersoll–Ross(CIR) process Cox et al. (2005) with the drift function depending on the day-of-week, and the daily discharges $d(t)$ come from simulating patients movements within hospital units. All the parameters to simulate the arrival and discharges are calculated empirically using real data. We provide an overview of the CIR model, a description of the real dataset, and details of the semi-synthetic generation in the rest of this section.

B.1. Cox–Ingersoll–Ross model

In generating the arrival process, we assume the arrival rates on different days are *random* and follow the CIR process. The standard CIR process can be characterized by the following SDE:

$$dr(t) = \alpha(\mu - r(t))dt + \sigma\sqrt{r(t)}dW(t) \tag{22}$$

where W_t is the Wiener process, μ represents the long-term mean, α represents the speed of the adjustment to the long-term mean, and σ represents the variation of the process. Note that the drift function, $\alpha(\mu - r(t))$, in the standard CIR process is time-stationary. However, the real data shows the hospital arrivals exhibit a strong day-of-week pattern. We describe how we modify the standard CIR process to have a time-varying drift function in Section B.3.

To simulate arrivals from the CIR model, one common approach is through the Euler–Maruyama method, which provides an approximated numerical solution:

$$r(t) = \max\Big(r(t-1) + \alpha(\mu - r(t-1))\Delta t + \sigma\sqrt{|r(t-1)|}\sqrt{\Delta t}z_t, 0\Big), \tag{23}$$

where the process uses $\max(\cdot, 0)$ to ensure that there are no negative values appearing during the approximation, which is one of the properties in the CIR model.

B.2. Description of the real dataset from paterner hospital

The real dataset comes from a partner hospital in the state of Indiana. The dataset contains patient-level movement history between different units in the hospital. The data spans from 2020 to 2021. The units can be categorized into two types: Medical/Surgical units (non-ICU units) and ICU units. For each patient, the

data contains timestamps on their arrival time to each unit, the transfer-in/out times between units, and the discharge time from the hospital. With these time stamps, we can estimate the empirical daily arrival rates for the two types of units and the length-of-stay distributions in each type of unit.

We use the following notations for these estimated quantities. For each day $a_{hos,t} = \sum_u a_{u,t}$ denotes the total arrival rate on the day t and $a_{u,t}$ the arrival rate to units u, where $u \in U = \{nonICU, ICU\}$ denotes one of the two types of units. Assuming we have $T = 7n$ days in total with n samples for each day-of-week, then

- *mean of arrival rate by day-of-week*: $\{\mu_1, \ldots, \mu_7\}$, where $\mu_i = 1/n \sum_{w=0}^{n} (a_{hos,i+7w})$;
- *standard deviation for arrival rate by day-of-week*: $\sigma_i = 1/n \sum_{w=0}^{n} (a_{hos,i+7w} - \mu_i)$;
- *routing probability*: $p_u = 1/T \sum_{t=1}^{T} \frac{a_{u,t}}{a_{hos,t}}$ for each u;
- *LOS distribution*: $p_{u,s}^{dis} = \frac{X_{u,s}}{X_u}$;

where $X_{u,s}$ denotes the number of patients stayed in unit category u for s days and X_u denotes the total number of patients stayed in this unit category. For the LOS distribution, we further assume that the maximum LOS is 4 days (validated by the data as the proportion of patients staying longer than 4 days is small). With these parameters estimated empirically from the real dataset, we then use them to generate semi-synthetic data as described in Section B.3.

B.3. Semi-synthetic data generation

Algorithm 1 Semi-synthetic Data Generation

Generate Arrivals: First, we generate the arrivals by the numerical CIR process with the parameters $\{\mu_1, \ldots, \mu_7\}$ depending on the day-of-week,

$$a(t) = \max\Big(a(t-1) + \alpha_t\big(\mu_{t\%7} - a(t-1)\big)\Delta t + \sigma_{t\%7}\sqrt{|a(t-1)|}\sqrt{\Delta t} z_t, 0\Big)$$

Assign Arrivals to Units: Here we have two units as $U \in \{MS, ICU\}$:

$$a_u(t) = Bionomial(a(t), p_u), \quad \text{for } u \in \{nonICU, ICU\}$$

Generate Discharges: Then, we can generate the number of discharges in the 4 future days according to the length of stay probability table:

$$\tilde{d}_u(t+i) = Multinomial(a_u(t), p_{u,i}^{dis}), \quad \text{for } i \in \{0, 1, \ldots, 4\}$$
$$d_u(t) = \sum_{j=t-4}^{j} \tilde{d}_u(j)$$

Generate Census: Since we have generated each day's arrivals and discharges, we can generate our census:

$$x_u(t) = x_u(t-1) + a_u(t) - d_u(t)$$

Algorithm 1 describes the procedure of generating the semi-synthetic data. Note that to capture the day-of-week pattern in the arrival rates, we modify the standard CIR process to have a time-varying drift function, where μ_i follows a periodic pattern with one week (7 days) as the period. Correspondingly, we need to adjust the mean reversion factor α_t to be time-varying through a weekly update scheme, e.g., $\alpha_1, \ldots, \alpha_7 = 0.1$ and $\alpha_8, \ldots, \alpha_{14} = 0.2$. We let α_t gradually increase to one during the first five weeks to capture the transient effect. The primary benefit of this semi-synthetic generation via Algorithm 1 is that it allows us to calculate the "ground truth" parameters. For example, the expected number of arrivals, discharges, daily census, etc. With these calculated numbers, we could compare them with corresponding results estimated from the generative models for evaluation.

Appendix C: Enlarged Figure for Delta Coverage Network Design

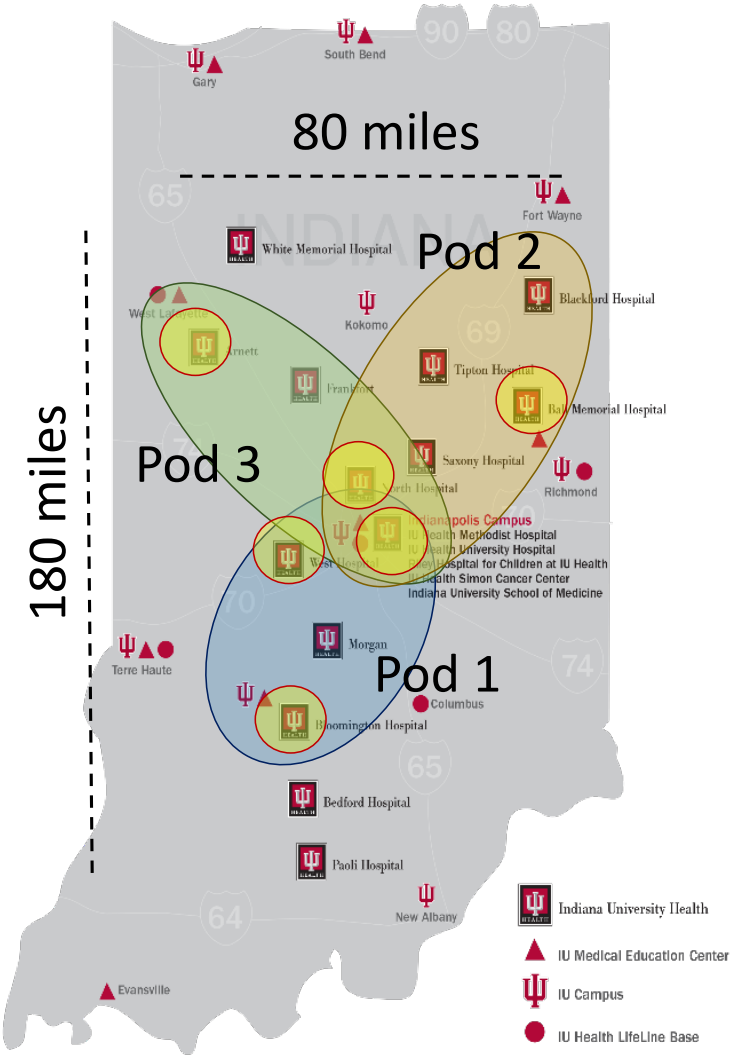


Figure 7 **Final network configuration for the Delta Coverage program. Small yellow circles are participating hospitals. The larger non-yellow circles are the pods of hospitals, each with their own Delta Coverage team. A Delta Coverage team only floats within their own pod.**

Appendix D: Delta Coverage Dashboard Functionality and Features

In this section, we describe the Delta Coverage Dashboard and how it supports on-call and deployment decision making in a variety of ways.

Dashboard Functionality and Usage. Once or twice a day, non-DC staffing data for the next 21 days is pulled from the Kronos time-keeping database and from a separate DC staffing database. The latter was manually curated in order for the DC program's implementation team to have more control over the datastream as the program was being rolled out. We also pull patient location data from the enterprise data warehouse which provides information about individual patient movement for the last 30 days. The movement data contains the location of each patient on each hour of the day. The granular patient location is then sent through a data-pipeline where it is cleaned of (significant) data-errors and converted into daily patient arrival rates (emergency department or elective admission), discharge rates, and occupancy acuity level (M/S, PCU, ICU) at each hospital. The data is gathered separately for day vs night shift, with day shift data coming from 11AM and night shift data coming from 11PM.

Once the data passes through the pipeline, it is fed into the prediction model that can generate census sample paths to input into the stochastic optimization model. The optimization is the run using a warm-start approach where the algorithm uses the previous day's optimization solutions to most efficiently use the computational power allocated to the pilot. The on-call and deployment decisions along with the current staffing plan and expected nurse demand at each hospital are written to a platform-agnostic csv file. The output data-file is then read into a user interface that the Delta Coverage design-team created to inform Delta Coverage scheduling and deployment decisions shown in Figure 8. The left panel of the figure allows the user to display different views of the data in graphical form (e.g., bottom right panel that plots the nurse demand vs the staffing) or table form by allowing the user to filter the table based on the selected criteria. The user can select day or night shift (upper left panel), any subset of hospitals (one panel down), deployment group that denotes which set of DC nurses being considered for transfer.

The DC nurse manager deploys DC nurses scheduled for the current day based the optimization model's deployment (recourse) decisions that are fed into the Delta Coverage Dashboard. Once a week, the DC manager informs the DC nurses of their planned work location (on-call location) based on the optimal on-call decisions coming from the most recent run of the full optimization model.

Dashboard Visualization Features. One of the key features of the Delta Coverage Analytics application that both supports user decision making, design for adoptability, and the change management process is a suite of visualizations to help users understand the impact of the nurse deployment actions on the broader system. The visualizations allow for

(i) Heatmaps detailing the level of understaffing at all the different hospitals before and after the deployment decisions;

(ii) Graphs of the past and forecasted occupancies and the nursing staff utilization before and after deployment decisions;.

These features are integral to the Delta Coverage decision and execution process as they allow

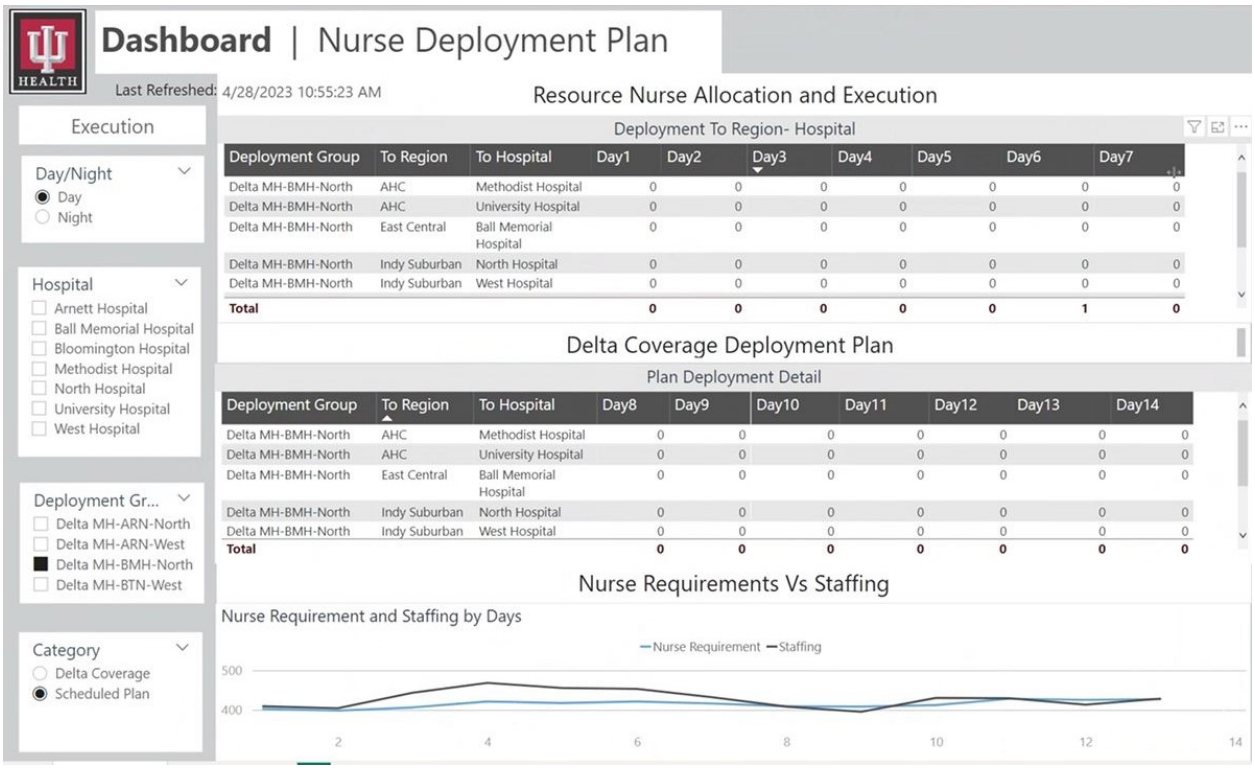


Figure 8 Full view of Delta Coverage Dashboard front page.

(i) Users to test what if scenarios and get immediate feedback on how changing the optimal recommendations would impact the system

(ii) Management to provide evidence to the individual hospitals of WHY the decisions are being made and how the decisions increase the fairness in the system.

As an example, the bottom right panel of Figure 8 displays the forecasted demand and scheduled nurses over the next two weeks. In the bottom left panel of the figure, users can adjust which staffing plan to view. On the main page they can view the staffing and demand based on schedule as it stands or the recommended schedule after the optimal deployment decisions (before and after). This way managers can immediately see the impact of the optimization recommendation.

Appendix E: Detailed Post-pilot Analysis

We had three phases of performances that tie to the three phases of implementation. In the pre-implementation part (historical counterfactual), the two month analysis suggested a 4% reduction in understaffing. In phase 1, the live testing of the analytics suite, we projected based on the 5 month test window that the Delta Coverage program could potentially reducing understaffing by 5% and overstaffing by 1%.

The phase 3 performance analysis was the most critical, given that it was based on the full pilot implementation where we were able to learn exactly how the analytics suite could be used in combination with additional knowledge of nurse managers using the dashboard. During this phase we expanded our performance metrics to include equity among Delta Coverage nurses and equity among hospitals participating in the pilot; definitions of equity are detailed in the associated subsection.

To perform the analysis, we compare two cases for a fair "apples to apples" assessment of the pilot program.

Case 1: We (counterfactually) assign each of the Delta Coverage nurses to a fixed hospital location ("home hospital") and do not allow them to work at any other hospital (standard pre-Delta Coverage approach).

Case 2: Is what was actually implemented in the pilot (which involves moving nurses based on the Delta Coverage Analytics Tool)

For each shift that was worked by a Delta Coverage nurse, we compared the actual understaffing (nurses required minus nurses working with zero being the lower bound) that occurred with what the understaffing would have been had the Delta Coverage nurse worked the shift in their (counterfactual) "home hospital." We utilize the same method for overstaffing. Next, we present the impact of Delta Coverage on the system as a whole by calculating under- and over-staffing metrics across all Delta Coverage shifts in all participating hospital E.1.

E.1. System Level Metrics

We were pleased to discover that the results of the pilot from May 7 to June 23, 2023 were better than our initial dry run projected. In this analysis we consider the impact that the Delta Coverage program has had on understaffing and overstaffing in terms of number of understaffed shifts eliminated, percent reduction in understaffing, and estimated annual cost savings from the program.

Understaffing. Among the shifts that the DC nurses worked, in a little more than 1 month (36 days) the Delta Coverage pilot has

- reduced the number of understaffed shifts by 24.5 (in 36 days). This is equivalent to
  - 248 fewer understaffed shifts per year (25 shifts per DC nurse per year)
  - reducing understaffing by 13%
  - 1 fewer understaffed shift for every 4 shifts worked by a Delta Coverage nurse
- Reduced the number of overstaffed shifts by 28.5 in 36 days. This is equivalent to
  - 289 fewer overstaffed shifts (29 shifts per DC nurse per year)
  - reducing overstaffing by 5%
  - 1.2 fewer overstaffed shifts for every 4 shifts worked by Delta Coverage nurse
- $400K in cost savings
  - Cost avoidance was calculated based on overtime costs associated with understaffing, premium pay for extra nurses that are not needed, overstaffing costs, among other metrics.
- Lives saved, nurses retained, mistakes avoided, better care given.

E.2. Delta Coverage Nurse Work Variety, Stability, and Equity.

To measure equity in terms of how Delta Coverage nurses are used in the program, we measure the proportion of time (shifts) each nurses spends at a remote facility. Of interest is that (1) each Delta Coverage nurse have sufficient variety of working location based on the feedback that these nurses want to travel (which is why they joined the program) but also want some stability of working in their home hospital, and (2) comparing

between nurses, each Delta Coverage nurse should have a similar amount of variety in the working location so that the travel regime is fair across Delta Coverage nurses.

Figure 9 provides a high-level visual summary of Delta Coverage nurses' work schedules. For the 10 individual nurses participating the 6-week pilot, Figure 9 shows the percentage of shifts worked at different hospitals for each of them. Some nurses worked in a pod of three hospitals and others worked in a pod of two hospitals. In general we see a pattern that shows that the nurses have fairly similar distributions of

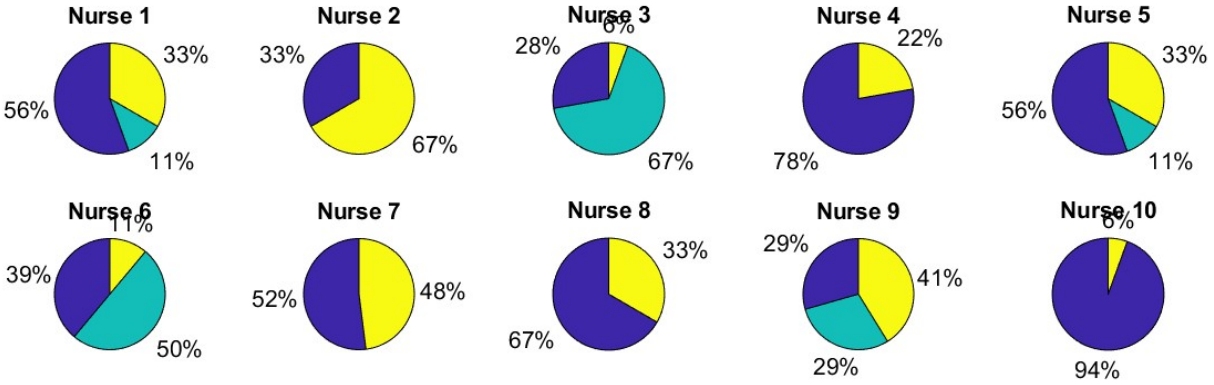


Figure 9 Fraction of shifts worked at each of a nurses locations for all 10 Delta Coverage nurses.

work locations (compare nurses with 3 hospitals separately from nurses with two hospitals). Recall that we do not need the shifts to be evenly distributed among hospitals, but rather that all nurses have a similar distribution of shifts across hospitals. As a final note, Nurse 10 was certified in 1 (out of 3) acuity levels, which somewhat restricted their transfer capability.

Additionally, we capture (1) the **variety** of opportunity - are they visiting each hospital regularly enough to gain experience and earn travel premium and (2) the **stability** of each individuals' schedule from week to week, and (3) the **equity** among Delta Coverage nurses for measures (1) and (2). These metrics were summarized in Table 1 in the main paper. We now explain more details about the calculation of these metrics. To measure the work variety and equity, we use the *Gini coefficient*, which is commonly used as a measure of dispersion in many fields. The Gini coefficient lies between zero and one, with zero representing perfect equality and one being perfect inequality. In our context, a Gini coefficient of zero in terms of work variety means that the nurse spends an equal amount of time at each of the hospitals in their catchment. Similarly if they spent all their time in one hospital then the Gini coefficient would be one. We do not set a target on work variety, but rather a target such that all the nurses have similar work variety, since traveling to different hospitals is the only difference between a DC nurse and a resource nurse.

When discussing equity in the subsequent paragraphs, a general rule of thumb is that a Gini coefficient of 0.3-0.4 is considered fair and 0.2-0.3 is considered very fair. With respect to equity between nurses, a smaller Gini coefficient means that individual nurse's work variety / schedule stability is close to each other, indicating a fair implementation of the program. We use this interpretation of the Gini to evaluate our

metrics as well. To measure stability, we use the coefficient of variation of each nurse's work variety over time calculated over the weekly the Gini for each nurse.

Work Variety and Equity. Work variety is measured at the individual level by obtaining one Gini coefficient for each individual for the measurement period (May-June). The average work variety (mean of the Gini coefficient) across all Delta Coverage nurses is 0.42. Note that Nurse 10 was an outlier due to lower flexibility in what roles they could fill so we remove the outlier when calculating equity in work variety. After doing so, the equity in work variety measured across Delta Coverage nurses has a Gini coefficient of 0.3, which is very fair.

Schedule Stability and Equity. To measure the stability of a Delta Coverage nurse's schedule, we calculate the variability in work variety from week to week. Quantitatively, for each nurse we first calculate work variety for each week using the Gini method above. Next, for each nurse we calculate the coefficient of variation (CV) of their work variety over the 6 week horizon of the pilot. The CV is the standard deviation of work variety over the course of the pilot divided by the mean, which is a common normalized measure of variability. The smaller the CV, the less variable the nurse's work variety. We adopt the convention that $CV < 1$ is consider low variability and $CV > 1$ would be high variability. Considering all nurses, the average CV of work variety is 0.41 and the equity (Gini of the CV) is 0.31, indicating that the program is creating schedules that are both stable and very consistent / fair across Delta Coverage nurses.

Delta Coverage Hospital Equity. To measure fairness of the allocation of Delta Coverage nurses to hospitals, we again use the Gini coefficient. After removing the single outlier (BTN) mentioned in the introduction, which maintains the concept of fairness since that hospital was well-staffed during the pilot period, the Gini coefficient was 0.29 and indicates a very fair allocation.

Recap. In summary, the previous analyses have demonstrated the the pilot has not only achieved significant reductions in under- and over-staffing, but is also creating nurse schedules and allocating Delta Coverage resources in a desirable and equitable manner.