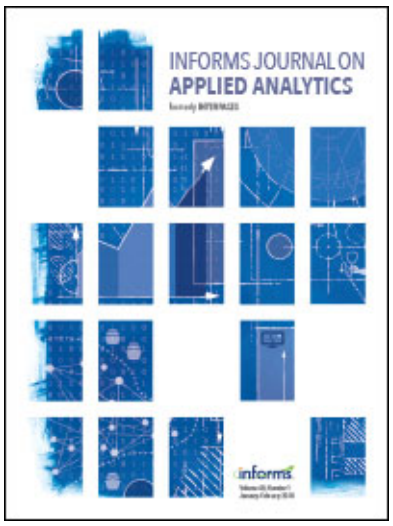


INFORMS Journal on Applied Analytics

Publication details, including instructions for authors and subscription information: http://pubsonline.informs.org

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

Jonathan E. Helm, Pengyi Shi, Mary Drewes, Jacob Cecil

To cite this article:
Jonathan E. Helm, Pengyi Shi, Mary Drewes, Jacob Cecil (2024) Delta Coverage: The Analytics Journey to Implement a Novel Nurse Deployment Program. INFORMS Journal on Applied Analytics 54(5):431-454. https://doi.org/10.1287/inte.2024.0140

https://pubsonline.informs.org/journal/ijaa

INFORMS JOURNAL ON APPLIED ANALYTICS
Vol. 54, No. 5, September–October 2024, pp. 431–454
ISSN 2644-0865 (print), ISSN 2644-0873 (online)

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

Jonathan E. Helm,[a] Pengyi Shi,[b,*] Mary Drewes,[c] Jacob Cecil[c]

[a] Kelley School of Business, Indiana University, Bloomington, Indiana 47405; [b] Daniels School of Business, Purdue University, West Lafayette, Indiana 47907; [c] Nursing Organization, Indiana University Health, Indianapolis, Indiana 46202

*Corresponding author

Contact: helmj@iu.edu, https://orcid.org/0000-0001-5577-5530 (JEH); shi178@purdue.edu, https://orcid.org/0000-0003-0905-7858 (PS); mdrewes@iuhealth.org (MD); jcecil2@iuhealth.org (JC)

https://doi.org/10.1287/inte.2024.0140

Abstract. Amidst critical levels of nurse shortages, we partnered with Indiana University Health (IUH) to pioneer a novel suite of advanced data and decision analytics to support a new model of nurse staffing. This statewide 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 travel and resource nurses by adding flexibility to move nurses *between hospitals* to dynamically respond to short-term patient census fluctuations. This paradigm shift necessitated the development of analytics to execute these interhospital transfers. Specifically, we develop analytics to create a two-week advance on-call list for travel and a 24- to 48-hour call-in decision. Our Delta Coverage Analytics Suite was launched in October 2021 as a Microsoft PowerBI application and provides an integrated solution that has supported and continues to support this new staffing approach at a statewide scale. The suite contrasts with existing nurse scheduling tools that primarily cater to single hospitals or units. It incorporates (1) a novel patient census forecast based on a deep generative model capturing complex spatial-temporal correlations and avoiding error accumulation occurring in traditional time-series models and (2) a stochastic optimization that prescribes optimal on-call and deployment decisions. The pilot, conducted from May to June 2023, produced a remarkable reduction in understaffing, with estimated annual savings of $2.5 million to IUH and over $1.5 billion on a national scale compared with the conventional solution of hiring travel nurses. As the first program of its kind, our methods establish new benchmarks for evidence-based and data-driven nurse workforce management with the potential to transform how healthcare institutions approach the national nursing shortage crisis.

History: This paper has been accepted for the *INFORMS Journal on Applied Analytics* Special Issue—2023 Daniel H. Wagner Prize for Excellence in the Practice of Advanced Analytics and Operations Research.

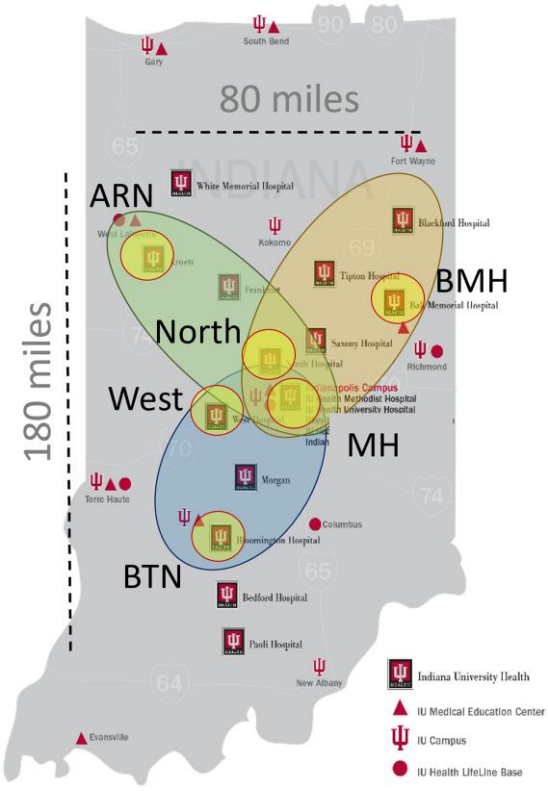
Keywords: nursing shortage crisis • nursing practice innovation • analytics for staffing • machine learning forecast • predictive-prescriptive integration

Introduction

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 by 2030 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 term and the long term. This paper presents one such innovation that was codeveloped and successfully implemented in partnership with Indiana University Health (IUH): the Delta Coverage (DC) internal travel nursing program. As the largest healthcare system in Indiana with 16 hospitals and over 9,000 nurses, IUH serves over 1.4 million residents across five diverse regions spanning 14,000 square miles. The DC program, to the best of our knowledge, is the first *implemented statewide program* that utilizes a flexible pool of full-time resource nurses capable of providing care in multiple hospitals and adjusting their work location on short notice in response to understaffing. In contrast to typical travel nursing arrangements with 12-week contracts, the DC program executes short-term deployments on the scale of days rather than months to dynamically respond to geographic and temporal fluctuations in hospital occupancies. Figure 1 shows the implemented DC network design and IUH's catchment area, which highlight its statewide coverage.

Our collaborative efforts led to the development of the Delta Coverage Analytics Suite, a comprehensive solution and implementation that leverages state-of-the-art predictive and prescriptive analytics, without

Figure 1. (Color online) The DC Network Design Consists of Three Pods and Spans 180 by 80 Miles



Notes. The squares are IUH hospitals, the circles indicate the six pilot hospitals, and the ellipses are DC pods. DC nurses can be deployed to any hospital within their pod. IU, Indiana University.

which the DC program would not have been feasible. The DC analytics suite dynamically optimizes nurse deployment and staffing on a multiple-hospital scale in contrast with off-the-shelf nurse scheduling analytics, which usually target individual units or hospitals and do not require multiple-day advanced notice prior to reassigning nurses. The distinctive dynamics and complexities of real-time nurse deployment over a large network make it difficult for existing solutions to gain traction in the nursing market, presenting an opportunity for our DC analytics suite to make a significant step forward in addressing the nurse staffing crisis.

Implementation and Impact

Launched in October 2021, the analytics suite underwent three phases of implementation. The upper panel of Table 1 summarizes key performance indicators extrapolated to annual estimates from the pilot (the last phase), which ran from May to June 2023 for six weeks. The left half of this upper panel ("Direct impact") shows the overall impact of the DC program versus a counterfactual that mimics hiring the same number of non-DC nurses (non-DC), such as travel nurses. The right half shows the "Marginal impact" (additional benefit) of the DC program over the non-DC counterfactual. Our pilot showed significant results: a 17% reduction in understaffing, equating to a projected 340 fewer incidents of understaffed shifts annually. This was made possible by moving 10 DC nurses among six hospitals participating in this initial pilot. This compares with a 4% reduction in understaffing (90 fewer shifts annually) when hiring 10 non-DC nurses. That is, for each understaffed shift eliminated by a non-DC nurse, a DC nurse can mitigate 340/90 = 3.7 understaffed shifts.

Further analysis of the pilot indicates significant increases in nursing productivity: to achieve the same understaffing mitigation as provided by the 10 DC nurses, IUH would have needed to hire *at least* 16–19

Table 1. The Performance of the Delta Coverage Pilot from May 2023 to June 2023

	Direct impact		Marginal impact	
Reduction	Understaff (DC)	Understaff (non-DC)	Understaff vs. non-DC	Overstaff vs. non-DC
Annualized shifts	340	90	250	290
Improvement, %	17	4	13	43
	Work variety (Gini)	Sched stability (CV)		Hospital DC shifts used, %
Average	0.36	0.3[a]		19
Equity	0.3[b]	0.31		0.29[b]

Notes. The upper panel shows the system-wide value of the DC program. We compare the annual number and percentage reduction in overstaffing and understaffing with the scenario of hiring the same number of non-DC nurses. The lower panel shows the average value and equity score across all DC nurses and hospitals. "Work variety" and "Hospital DC shifts used" are measured by the Gini coefficient. "Schedule stability" is measured via the coefficient of variation (CV). See Appendix F for details of the calculation of these metrics.

[a]A smaller value is better, with ≤ 0.5 being very stable.

[b]A value of ≤ 0.3 is generally considered very equitable.

non-DC (either regular or traveler) nurses over the six-week horizon, assuming that IUH staff could predict precisely when and where understaffing would occur in each hospital over the six-week period. In a more realistic scenario, this estimate could rise to hiring 19 non-DC nurses; see Appendix F.1 for details. Therefore, one DC nurse is equivalent to 1.9 non-DC nurses in addressing understaffing.

This significant productivity gain is attributed to the flexibility of DC nurses who can be deployed to different hospitals, whereas non-DC nurses must be hired for a specific hospital. To illustrate, consider a scenario in which one hospital experiences understaffing during the first half of a week but in which another hospital faces understaffing during the second half. A single DC nurse can cover both hospitals, whereas the traditional approach would require hiring two additional nurses, one for each hospital. Meanwhile, the hiring of 10 DC nurses led to 43% fewer overstaffed shifts compared with hiring 10 non-DC nurses (290 fewer shifts per year). Note that the improvement would be even more significant if we accounted for the rigid 12-week contracts of travel nurses. For example, if a hospital experiences understaffing for only the first six weeks, the contract cannot be canceled, leading to overstaffing in the remaining six weeks. In other words, the DC program achieves "pooling" effects both geographically and temporally. In addition to the efficiencies generated by the DC program, the final allocation of DC nurses is considered to be fair to participating hospitals and nurses. See the lower panel of Table 1 for a summary, and see additional discussion in the Equity and Adoptability section.

The deliberate decision to limit the initial pilot's scope stemmed from the inherent uncertainty associated with this new staffing method. The size of the pilot, 10 DC nurses, was chosen based on the budget allocated to the pilot. Hiring 10 new nurses requires significant expenditure, and management determined that 10 hires provided an appropriate balance of proof of value and risk. Although the 10-nurse pilot may seem modest in scale, it was instrumental in validating our approach. Moreover, after this pilot's success, we are actively expanding the program to include the entire cohort of 300 resource nurses at IUH in the DC program. Our analytics suite has undergone comprehensive testing, and it is fully prepared to operate at this more significant scale.

Paradigm Shift

Historically, the idea of relocating nurses between hospitals has encountered skepticism and substantial logistical and cultural barriers. Although resource pooling is a well-known concept for improving efficiency in various industries, applying it to highly skilled medical professionals is a far more complex endeavor than pooling products and materials. At first glance, our pilot, which moves 10 nurses between hospitals, may appear to be a modest step. However, it represents a reshaping of traditional staffing paradigms that rely heavily on travel nurses. The core value of this pilot is its role as a proof of concept of an innovative solution that has the potential to revolutionize nursing practices and address a global crisis.

To elaborate, the DC program and analytics suite provides an alternative to the conventional response to shortages: hiring costly travel nurses. In contrast to DC nurses who can move between hospitals daily, travel nurses are typically hired on 12-week contracts with the same hospital and hence, provide at most the value of a traditional resource nurse. Although travel nurses can technically move to different hospitals after 12 weeks, they are unable to respond to the short-term fluctuations in nurse demand (one day to 3 weeks) that the DC nurses are designed to cover. In addition, if a travel nurse is not needed for the entire 12 weeks, that nurse still must stay on staff, resulting in unnecessary costs or sometimes having to send one less expensive, full-time nurse home when the hospital is overstaffed.

In contrast, our pilot demonstrates the feasibility of relocating nurses between hospitals without causing disruptions in hospital culture. Unlike travel nurses, who often lack familiarity with hospital teams and processes, DC nurses are IUH employees and thus, are part of the culture, seamlessly integrating into care teams across multiple hospitals. More importantly, the DC program delivers significant cost savings because of the lower cost of DC nurses and the "pooling" effects; hiring 10 DC nurses costs approximately $10 \times \$2,698 = \$26,980$ per week based on an estimated 75% hire compensation versus a full-time unit-based nurse. Hiring 10 DC nurses is equivalent to hiring 19 travel nurses as discussed. Travel nurse salaries have easily exceeded $4,000 per week since the pandemic. Therefore, 19 travel nurses would cost IUH at least $76,000 per week. This corresponds to $2.5 million annual savings, even at the pilot scale.

Staffing costs have always been a significant portion of hospital budgets, and these expenses have surged, particularly since the pandemic, with travel nursing being a major cost driver (American Hospital Association 2022). This unsustainable financial strain will eventually translate into higher cost burdens for patients, the healthcare system, and taxpayers. Within this context, the potential impact of our DC program is profound; more than 4,000 hospitals (67%) in the United States are associated with 626 health systems (Furukawa et al. 2020). If each of the 600+ health systems was to employ 10 DC nurses and the DC analytics suite, the national impact would exceed $1.5 billion dollars annually. Therefore, the significance of the DC implementation extends far beyond our initial pilot. Its success marks a

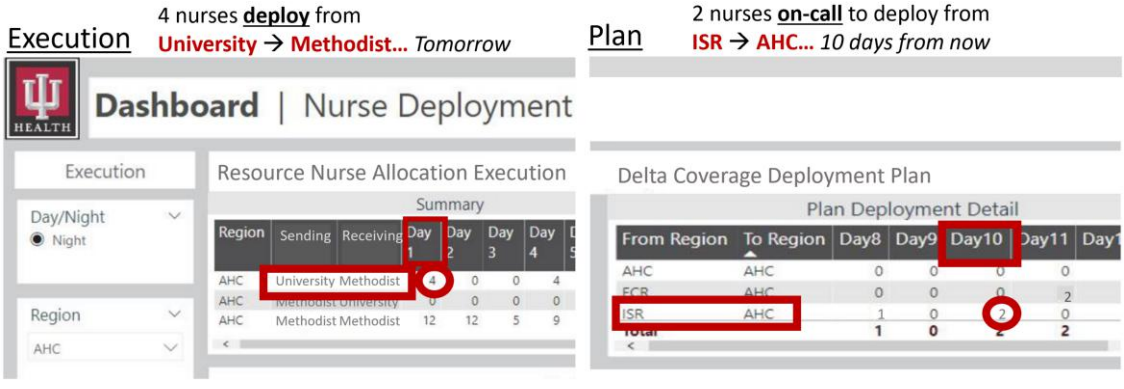
turning point in the way that healthcare institutions manage their nursing workforce, moving beyond the inefficient yet previously unavoidable practice of heavy reliance on travel nurses. Unlike innovations in inventory or supply chain management, which deal with goods or machines, our experiment is distinctly human centered. The complex nature of the healthcare industry, coupled with its deeply ingrained resistance to change, made convincing a hospital system to embrace this new practice initially seem nearly impossible. Now, this demonstration of value serves as an invitation for other healthcare providers to adopt similarly innovative solutions to meet the urgent demands of the healthcare and nursing industries. Notably, the achievement resulting from our collaboration with IUH nursing is not merely an enhancement of existing practices through analytics; rather, it leverages analytics to create an entirely new approach to nurse staffing that expands the boundaries of traditional practice.

Criticality of Operations Research Support

The concept behind Delta Coverage is to allow highly skilled nurses to float and work in multiple units, including units in other hospitals in the network. The ultimate goal is to move the *right number* of nurses to the *right unit* at the *right time* in order to respond rapidly to fluctuations in staff and occupancy across hospitals. Unlike programs for traditional resource nurses, who usually float between units within an individual hospital and receive their assignments less than 24 hours before a shift, Delta Coverage requires sophisticated advanced planning that utilizes (1) predictive analytics to forecast occupancies for all 16 IUH hospitals and (2) prescriptive analytics to determine optimal on-call and call-in decisions for DC nurse transfers. To meet this critical 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 (1) and a new generative model-based stochastic optimization (SO) for component (2). Figure 2 provides a close-up view of the decision support for the two stages of decisions, with the "Plan" (the right panel of Figure 2) indicating how many nurses should be put on call to travel one to two weeks in advance (for example, from the Indianapolis Suburban Region to the Academic Health Center region in 10 days) and the "Execution" (the left panel of Figure 2) showing how many nurses should be called in for travel 24–48 hours in advance (for example, from Methodist Hospital to University Hospital the next day).

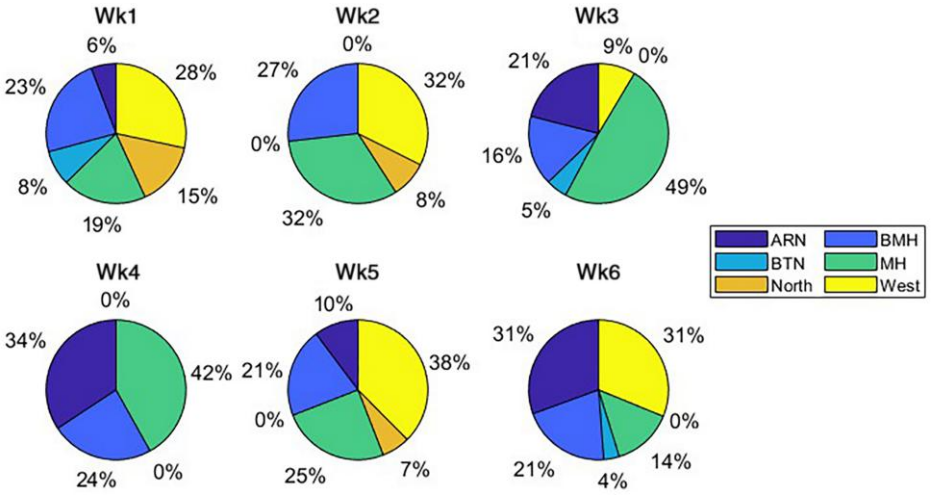
Our pilot program underscores the critical role of our operations research (OR)-based analytical solution in ensuring the success of this innovative practice, especially in a setting where this approach significantly departs from traditional practices and initially faced skepticism and resistance. Our analysis demonstrates that without following the prescriptive guidelines provided by the OR solution, the potential benefits would be significantly diminished. Before the pilot started, DC nurse movements were initially managed without the DC analytics suite for a few weeks. During this trial period, the DC program reduced understaffing by only 1.2% and overstaffing by 0.15%. In contrast, if the corresponding decision support system (DSS) recommendations had been followed (extracted from the back-end database), understaffing could have been reduced by 9.4%, and overstaffing could have been reduced by 2.4%. Without the support of the analytics suite, the entire innovation could have potentially been jeopardized because of marginal performance. The OR-driven decision-making process, rooted in data-driven insights, is the cornerstone of our program's success. It not only enables efficient nurse deployment but also optimizes resource allocation. This exemplifies the significant value of combining advanced analytics with OR to address pressing challenges.

Figure 2. (Color online) The Graphic Shows a Snapshot of the DC Dashboard Decision Support



Note. ISR, Indianapolis Suburban Region.

Figure 3. (Color online) The Pie Charts Show the Fraction of DC Shifts Allocated to Each Hospital by Week Weighted by Hospital Size



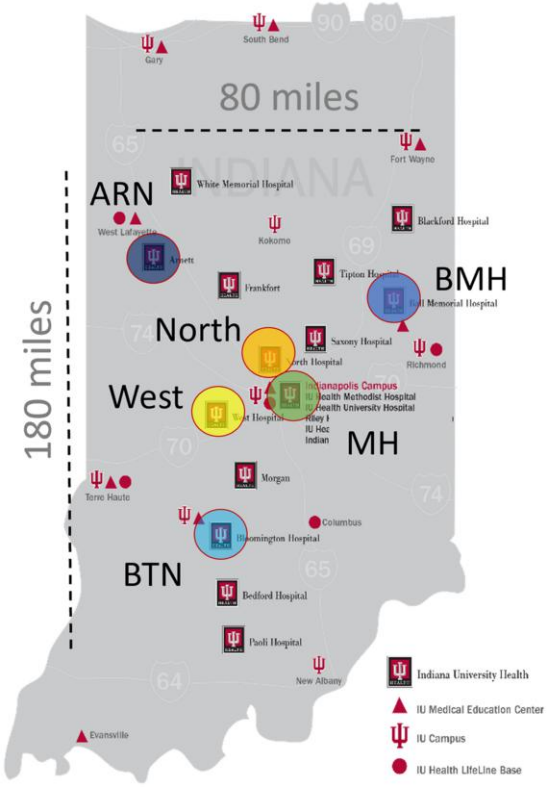
Equity and Adoptability

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 (Blegen et al. 2011, Aiken et al. 2014). 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 because of 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 nationwide that are facing similar challenges. In particular, one significant concern voiced by chief nursing officers (CNOs) of some IUH 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 any resources back. However, the implementation shows promising results for the hospitals located in more rural communities. Figure 3 provides a visual representation of the distribution of Delta Coverage resources among participating hospitals (shown in the map in Figure 4). Figure 3 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 Arnett Hospital and Ball Memorial Hospital, the two most rural hospitals in the pilot. See Appendix F for a comprehensive analysis of the pilot program's performance.

To summarize, the success of our pilot highlights the feasibility and benefits of internal travel nurse programs as an alternative solution for managing nurse shortages and optimizing workforce allocation, instead of solely relying on travel nursing. It introduces a new paradigm characterized by data-driven, analytics-

Figure 4. (Color online) The Graphic Shows a Map of the DC Hospitals



Note. MH, Methodist Hospital.

based decision making. It also has the potential for a far-reaching impact in the long run. This approach promotes workforce stability and a supportive environment, resulting in a more resilient and satisfied nursing workforce. Moreover, our analysis shows that the DC program's benefits extend to rural and marginalized areas that often bear the brunt of nursing shortages (because rural hospitals face more challenges in attracting and retaining nurses because of their remote locations), disproportionately affecting access to quality healthcare and population health outcomes in these areas. Our solution can effectively enhance 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 of development and implementation. In the section Delta Coverage Analytics Suite Details and Challenges, we present an overview of our outline, the challenges encountered, and our main contributions, which provide the road map for subsequent sections. To overcome the technical challenges, we first describe the novel multiple-hospital and multiple-unit nursing demand forecast based on a deep generative model in the section Generative Modeling to Predict Correlated Hospital Occupancies. We then introduce in the section Stochastic Optimization for Network Decision Making the prescriptive framework based on the stochastic optimization. In the section Integration of Predictive and Prescriptive Components, we discuss the seamless integration of forecast and optimization; the generative model structure perfectly complements our quasi-Monte Carlo (quasi-MC) 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 the section Implementation of Delta Coverage and Practical Challenges, we discuss the journey to launch the pilot implementation, including our tiered approach to build trust for deploying OR analytics for operational improvement. We conclude this paper with ongoing work in the Conclusion section.

Delta Coverage Analytics Suite Details and Challenges

Our analytics suite was implemented in three phases from October 2021 to June 2023 as a Microsoft PowerBI application: (1) live testing from October 2021 to April 2022, (2) program redesign and refinement with the leadership team from May 2022 to April 2023, and (3) 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 suite runs the following procedures on a daily basis.

1. On Monday, based on the demand forecast, scheduled nurses at each hospital, and available Delta Coverage resource nurses, determine the on-call list for a one-week period two weeks in advance (21 days ahead).
2. Each day at 4 a.m., update the patient census data and forecasts, and determine actual deployment decisions for the following day (24 hours later).
3. Load output into the Microsoft PowerBI dashboard to support decision making. The results 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.

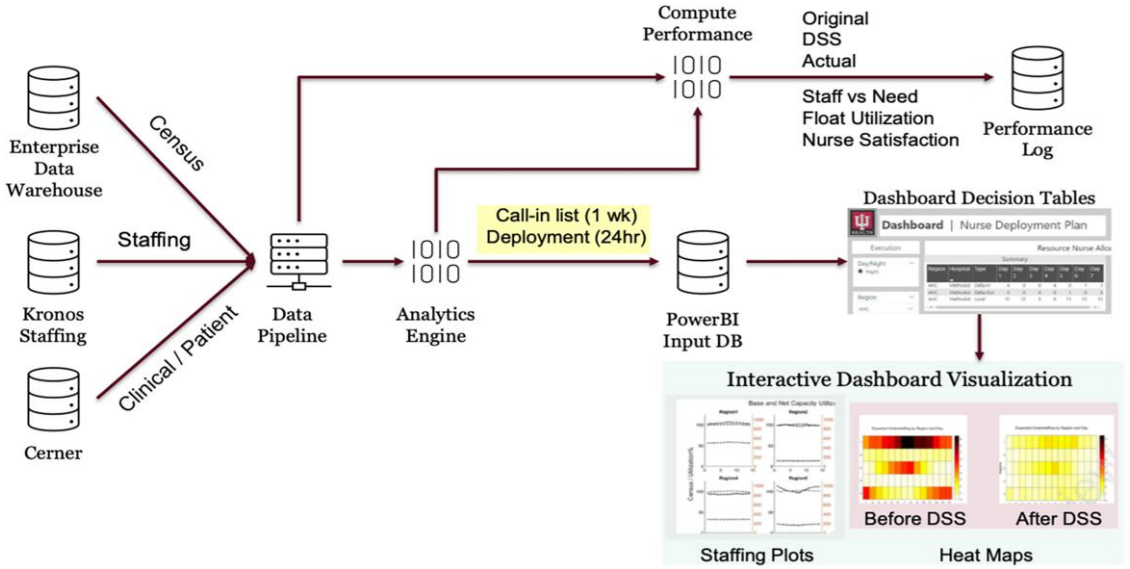
Figure 5 provides a schematic of the DC analytics suite design. The data required include the number of unit nurses and resource nurses scheduled at each hospital over the three-week planning horizon, the number of DC nurses scheduled for each day of the planning horizon, the current census at each hospital, and the history of patient movement over the past 30–60 days (to calculate arrival, discharge, and transfer rates that are used in creating the forecast). More details of the data and system functionality are provided in Appendix E.

Challenges

Given the goal of providing one to two weeks of notice to nurses who will be put on call to travel and 24 to 48 hours of notice on 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 because of 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 nonlinear correlations between hospitals. In our case, the decision structure that manages the transfer of nurses between hospitals complicates the system further, which

Figure 5. (Color online) The Schematic Details the Delta Coverage Decision Support Input Data and Workflow



Note. DB, database.

contrasts with 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 decisions critically depend on correlated census patterns over multiple days, and because (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 faced numerous practical obstacles. Penetrating the nursing industry with innovative practice and OR analytics has been exceptionally challenging as we discussed in the Introduction section. Convincing the industry to embrace a significantly different staffing model requires substantial evidence of its efficacy and benefits. Yet, off-the-shelf nurse scheduling analytics usually target individual units or hospitals, whereas 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 challenges during the implementation in Appendix F.

Literature Review

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 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 (VAEs) and generative adversarial networks (GANs), have the advantage of providing distributions as the output. Time-series generative models use GAN or VAE combined with recurrent neural network (RNN); for example, see Mogren (2016), Esteban et al. (2017), and Desai et al. (2021). TimeGAN (Timeseries Generative Adversarial Networks) (Yoon et al. 2019), considered as the current state-of-the-art method, combines autoregressive models with GANs and aligns the latent representations of real and generated data. However, these generative models have one primary limitation: learning stepwise conditional distributions that may accumulate errors and overlook key temporal patterns essential for downstream tasks; see more discussion in the section Generative Modeling to Predict Correlated Hospital Occupancies. Moreover, they often lack theoretical justification and interpretability, and they fail to consider the structural insights of realistic problems. In contrast, the predictive model that 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 (Medical/Surgical Unit) literature; for example, see Saville et al. (2019) and Griffiths et al.

(2020) for comprehensive reviews. Recent advances in analytics have helped to incorporate predictive analytics into nurse scheduling; for example, see Zlotnik et al. (2015), Ban and Rudin (2019), Spetz (2021), Anderson et al. (2022), and Shi et al. (2023). 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 by Hu et al. (2024), who used predicted patient demand to allow management to set base and surge staffing levels in an emergency department. It is important to highlight that this stream of literature has predominantly focused on staffing within individual or hospital units, which operate on a much smaller scale compared with 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 with patient or equipment transfer. For example, nurses need to move back to home locations after being transferred rather than being transferred again (in contrast to equipment that can be continuously moved). In addition, we need to design efficient and scalable algorithms to ensure practical implementations rather than treating the problem solely as a mathematical optimization problem.

Contributions

To the best of our knowledge, this work represents a novel implementation that leverages state-of-the-art predictive and prescriptive analytics to optimize nurse staffing in multiple hospitals and multiple units. Our focus is on a statewide program that dynamically reallocates nurses across a 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 temporal-based variational family based on patient-flow dynamics along with customized encoder-decoder structures for the learning. This both provides efficient representations of the census time series and generates distributional information for the decision optimization. Comparing our methodology with general-purpose prediction methods in the 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 more importantly, it provides a doubly stochastic patient census process structure for prescribing optimal decisions in the decision-support phase.
- Prescriptive innovation. We formulate an SO program to effectively capture essential trade-offs in our nurse deployment program while considering realistic implementation constraints. This SO integrates with the predictive model, making it unique in the sense that the demand is a doubly stochastic process in contrast with a conventional SO setup. This brings new computational challenges for sample-based methods because there are two layers of randomness. To efficiently solve the SO, we transform the original large-scale problem into a tractable linear program (LP) through a quasi-Monte Carlo method for scenario generation. At the heart of our technical innovation lies a seemingly complex modeling structure: doubly stochastic processes driven by multivariate Gaussian 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 presents a methodological contribution to the SO problem driven by doubly stochastic processes, which is understudied in the literature and may spark independent technical interest. Moreover, it presents a scalable solution that can be readily implemented by our partner.
- Implementation. Unlike prior research that focused on small-scale staffing optimization within individual units or hospitals, our work extends beyond those boundaries. We tackle the complex task of deploying nurses between hospitals using decision analytics across an entire state. As we discuss in the Paradigm Shift section, our pilot program serves as a proof of concept, demonstrating the feasibility and effectiveness of this innovative practice, which initially faced skepticism within the industry. Delta Coverage decision analytics offer an effective alternative to traditional travel nursing, thus making a significant contribution to the healthcare industry and broader society. Moreover, the implementation of decision analytics in the traditionally technology-resistant nursing industry represents a paradigm shift toward a more data-driven and evidence-based approach. This transition can foster a culture of continuous improvement and innovation, unlocking untapped potential and enabling informed decision making.

Generative Modeling to Predict Correlated Hospital Occupancies

To overcome challenges associated with existing time-series forecasts, such as the lack of distributional information and the 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. (2024), in which the authors developed a VAE method for temporal-based generative model learning. We tailor and adapt this framework to our specific hospital census prediction setting. Our adaptation captures the dependence structure and the time dynamics among the 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.

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 (i.e., the number of patients) on day t in k hospital units. We denote this time-series census sequence as $X_{0:T}$ for notational simplicity. 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. The pandemic is an example; the latent variables correspond to the disease spread and recovery, which drive the number of patients who will be hospitalized (arrivals) and how long they will need to be hospitalized (discharges). To capture this dependence, we use the generative modeling framework from Li et al. (2024) and tailor it to the hospital census setting. Specifically, starting with $X_0 = x_0$, the relationship of X_t, X_{t-1}, A_t, D_t follows

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

which captures the patient flow dynamics in hospitals—today's census equals yesterday's census plus arrivals and minus discharges—with some noise $\epsilon \sim N(0, \tau)$. 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). 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 (SDEs). As we elaborate, we do not directly learn the arrivals or discharges, and thus, we leave the specification of these SDE to Appendix B.

Cumulative-Difference Learning

A common way to learn the joint distribution of $\{X_t\}$ via the generative modeling framework is through stepwise learning: that is, learning the conditional distribution $X_t | X_{0:t-1}$ recursively for each day t. This method has 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. This is because in stepwise learning, the calculation of the current day's census is based on the previous day's census; for example, 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 adopt the 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, respectively (i.e., the net changes in X_t's). Then, we define a new variable that captures the cumulative difference: $\Gamma_t = X_t - X_0 = \sum_{i=1}^t \Delta_i = \sum_{i=1}^t (A_i - D_i)$. Here, Γ_t is the cumulative difference between the census on day t and the initial census $X_0 = x_0$. From Equation (1), the relationship between X_0, X_t, and Γ_t can be characterized as

$$X_t = X_0 + \Gamma_t + \epsilon_t, \quad \epsilon_t \sim N(0, \tau_t), \quad t = 1, \ldots, T. \tag{2}$$

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 captures the measurement errors, which are assumed to follow a multivariate normal distribution with zero mean and covariance τ_t. Note that $X_t \in \mathbb{R}^k$ is a multidimensional vector for the census in each of the k locations: hence, the covariance matrix $\tau_t \in \mathbb{R}^{k \times k}$. The covariance matrix is time varying because the noise ϵ_t for the cumulative difference changes over time.

Following the literature on generative models, we assume that the cumulative-difference sequence depends on the sequence of latent variables $\{Z_t\}$ through a set of stochastic difference equations $\Gamma_t = \Gamma_{t-1} + b_t(\Delta_{t-1}) + \sigma_t Z_t$, $t = 1, \ldots, T$ with the initial condition $\Gamma_0 = \Delta_0 = a_0 - d_0$. Here, $Z_0, \ldots, Z_k \sim N(0, I_d)$ are independent and identically distributed 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 SDE here can be seen as a discrete-time version of the Cox–Ingersoll–Ross (CIR) process.

VAE Learning Framework

To learn the unknown parameters, we maximize the log likelihood of joint distribution $p(\gamma_{1:T})$:

$$\log p_\theta(\gamma_{1:T}) = \log \int p_\theta(\gamma_{1:T} | z_{1:T}) p(z_{1:T}) dz_{1:T}, \tag{3}$$

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,

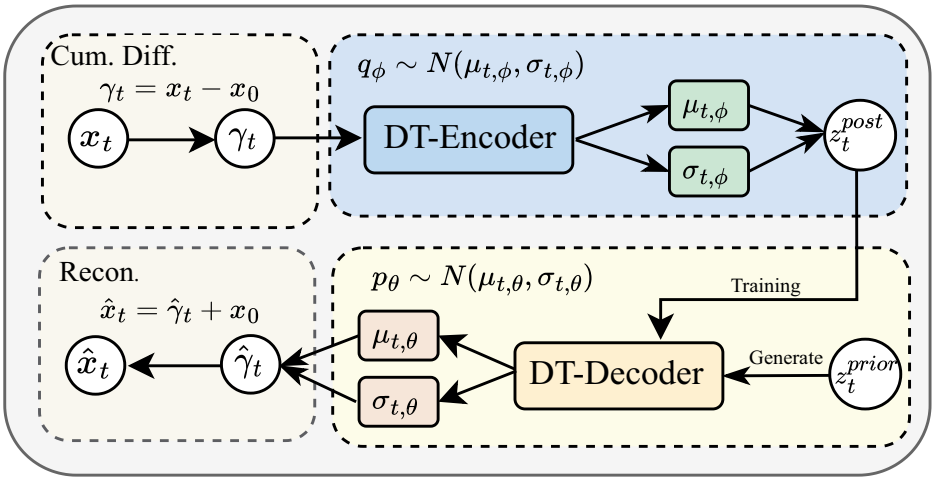
and θ represents parameters in the conditional distribution for $\gamma_{1:T}|z_{1:T}$. The likelihood function is intractable and hard to evaluate numerically. We adopt the VAE framework for the learning task. At a high level, VAE optimizes the evidence lower bound (ELBO) as the surrogate objective, which contains two major components: (1) learn the conditional distribution $p_\theta(\gamma_t|z_{1:t})$ via a *decoder* $f_\theta(\cdot)$ with parameter θ, and (2) 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 (1) is called the decoder because it decodes the latent variables $z_{1:t}$ to generate γ_t, whereas the variational distribution in part (2) is called the *encoder* because it encodes the 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 as well as the design of the encoder and decoders are relegated to Appendix B. Figure 6 characterizes the entire pipeline for the training and generation procedure. To summarize, we integrate domain knowledge by embedding the patient flow dynamics into the VAE framework. This allows our model to be interpretable, and importantly, it also provides a doubly stochastic patient census process structure for prescribing optimal decisions in the decision support phase.

Advantages over Other Machine Learning Models and Numerical Performance

In addition to the domain-aware design with specific patient-flow dynamics integrated within the learning, our prediction model offers two other advantages over conventional models. First, compared with traditional time-series forecast models, such as Autoregressive Integrated Moving Average, the encoder-decoder structure provides great flexibility to represent complex functional forms and allows for the easy addition of useful auxiliary covariates, such as the day-of-week or holiday indicators, to facilitate predictions. 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 difference learning setup and the mapping from $Z_{1:t}$ to Γ_t (captured via the decoder f_θ). Specifically, Γ_t is correlated with all previous $\Gamma_{1:t-1}$ because of 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.

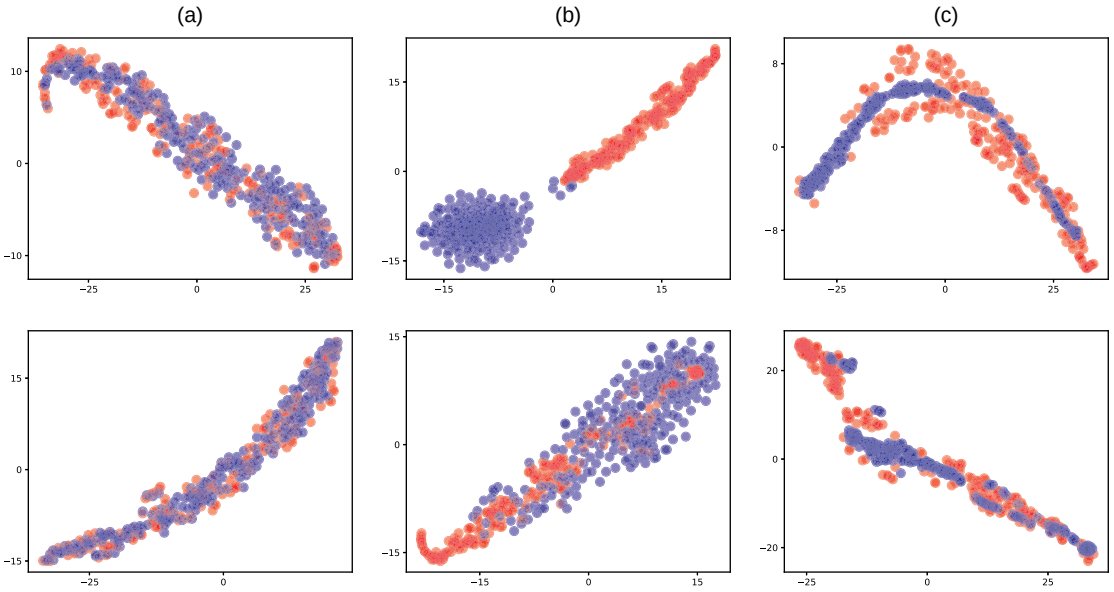
Second, by transforming the original census prediction problem into learning the cumulative difference, our method effectively avoids the error accumulation issue associated with recursive prediction that is commonly found in time-series generative models, including many state-of-the-art models, such as TimeGAN (Yoon et al. 2019). Because 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 stepwise learning. That is, the cumulative-difference 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 because it is solely determined by the latent variables. See Figure 7 for a comparison with benchmark algorithms, showing the advantage of our algorithm in addressing these issues.

Figure 6. (Color online) The Schematic Details of the Architecture of DT-VAE with Its Training and Generation Procedure



Notes. The DT-Encoder q_ϕ encodes input data to the latent space; the DT-Decoder p_θ generates data from encoder samples during training and a prior distribution during generation. Cum. Diff., cumulative difference; Recon., reconstruction; T-VAE, timeseries-variational auto encoder; t-SNE, t-distributed stochastic neighbor embedding.

Figure 7. (Color online) t-SNE Visualization for Our Algorithm, a Naive Time-Series VAE, and TimeGAN for Census Generation in Two Hospital Units



Notes. In each panel, the two sets of dots denote the original data and the generated data. Better mixing of the dots indicates higher-quality generated data. (a) DT-VAE. (b) T-VAE. (c) TimeGAN.

Stochastic Optimization for Network Decision Making

We built a two-stage stochastic optimization 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 optimization prescribes the weekly schedule on how many nurses to put on call for potential deployment one to two weeks in advance (step 1 of the DSS). Then, at the beginning of each day, based on the realized census and the updated forecast for the rest of the week, we again solve the optimization, and we use the first-day decision to determine the actual deployments on the current day (step 2 of the DSS).

In the optimization, the two levels of decisions are made sequentially. The first decision is the number of nurses to put on call each day for travel from their home hospital to a remote hospital. This decision is made prior to observing the census scenarios of the hospitals in the network. Then, after observing the census scenarios, the decision is made whether to deploy nurses who have been put on call to a remote hospital or to cancel the deployment so that the nurses will work their shifts in their home hospitals. If nurses are deployed to a remote hospital, they will work a minimum number of shifts at the remote hospital before returning to their home hospitals to avoid excessive travel. The secondment is an important design feature that ensures that a nurse does not have to travel two long-distance legs in addition to working a 12-hour shift.

The primary objective is to reduce system-wide understaffing without being too disruptive to nurses' lives through excessive or unreasonable travel schedules. To calculate understaffing, we account for the fact that nurse demand and the patient census are not equal. Instead, we calculate nurse demand by considering the patient-nurse ratios for different acuity levels; for example, one nurse is required for taking care of two patients in the intensive care unit (ICU) or four patients in the medical and surgical units.

In addition to understaffing costs, we consider other costs that can be tuned to achieve desired performance along multiple dimensions, such as the efficacy for the health system and attractiveness to DC nurses. These parameters include the following. The cost associated with the transfer decision results from two parts: (1) the fixed cost that compensates for the transfer and depends on the transfer distance and (2) the variable cost that compensates for premium pay during the length of the secondment. Tuning these transfer costs creates a system that has more or less churn: that is, the amount of travel that occurs across all DC nurses. If the costs are higher, then the system will generate less travel for the DC nurse pool on average; if the costs are lower, the system will generate more travel on average. If a transfer is cancelled during the deployment decision phase, we recoup a percentage of the transfer cost. This parameter determines how often a nurse who is

put on call will actually be deployed to a remote hospital. The lower the percentage of cost that can be recouped from the initial transfer decision, the less likely an on-call decision is to be canceled. Consequently, there will be a higher probability on average that a nurse will be deployed to an on-call destination. During program design, we adjusted these costs to achieve the desired system performance (see the Implementation of Delta Coverage and Practical Challenges section for additional details).

In addition, we utilized these tuning parameters along with constraints to achieve several design specifications, such as (1) limiting the number of times 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 an intermediate shift in their home hospitals, and (4) ensuring equitable use of Delta Coverage deployments to avoid perceived (or real) favoritism for certain hospitals. The full model specification is given in Table A.1 in Appendix A.

Integration of Predictive and Prescriptive Components

The most difficult task in evaluating the objective function of the stochastic optimization is the cost-to-go term, which is an expectation over all possible census scenarios. To evaluate this expectation, a common approach is to use the sample-average method. In our setting, the sampling-based optimization should fully account for the generative modeling structure used in the forecast step as specified in the section Generative Modeling to Predict Correlated Hospital Occupancies. That is, instead of directly sampling the census sequence X's as in conventional settings, we first sample the latent sequence Z's from a 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. In other words, although the two-stage SO developed in the Stochastic Optimization for Network Decision Making section may appear to be standard, it is different in the sense that the demand is a doubly stochastic process, contrasting with the conventional SO setup. This brings new computational challenges for sample-based methods because there are two layers of randomness. One of our main technical contributions in this paper is to develop an efficient algorithm, leveraging a quasi-Monte Carlo method and the special doubly stochastic structure, that efficiently overcomes this computational challenge. This could generate future technical research to study this type of new SO, which is uncommon in the literature and understudied.

To specify, recall that 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

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

where $N(0, 1)$ is a standard normal random variable. That is, X^i_t is a doubly stochastic random variable that depends on the latent variables $z_{1:t}$ and $\zeta_{i,t} \sim N(0,1)$. For the doubly stochastic random variable, 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 random variables $\zeta_{i,t}$'s. In the following, we let $\zeta = \{\zeta_{i,t}\}$ for the set of independent and identically distributed normal random variables 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 mth sample of the latent sequence, and let ζ^ℓ be the ℓth set of sampled random variables. In the interest of space, we focus on explaining the calculation of the understaffing part in the cost-to-go term. We define $y^{m,\ell}_{i,t}$ as the auxiliary variable that approximates the value of the understaffing function in unit i on day t given the mth sample $z^m_{1:t}$ and the ℓth sample ζ^ℓ:

$$\mathrm{Ex}\left[\sum_{t=1}^{T}\sum_{i=1}^{k}(X^i_t - \overline{n}^i_t)^+\right] \approx \frac{1}{M \cdot L}\sum_{m=1}^{M}\sum_{\ell=1}^{L}\sum_{t=1}^{T}\sum_{i=1}^{k} y^{m,\ell}_{i,t}, \tag{4}$$

s.t.

$$y^{m,\ell}_{i,t} \geq (x_0 + \mu^i_{t,\theta}(z^m_{1:t})) + \sigma^i_{t,\theta}(z^m_{1:t}) \cdot \zeta^\ell_{i,t} - \overline{n}^i_t,\ \forall i,t,\ell,m, \tag{5}$$

$$y^{m,\ell}_{i,t} \geq 0, \qquad \forall i,t,\ell,m. \tag{6}$$

Here, for ease of exposition, we suppress the dependence of $\overline{n}^i_t$ on the recourse decision, which can reduce the understaffing through the minimization in the second stage of the stochastic SO.

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 Equation (4). The benefit is that there is no correlation among these Gaussian random variables (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 (for example, $z_{1:T}$ has 21 dimensions when we plan for three weeks out with $T = 21$). The 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 because the open-source optimization solver cannot handle a large number of samples. To address this issue, we leverage the quasi-Monte Carlo method, which is known to reduce variance in sampling; it can improve the rate of convergence from $O(1/\sqrt{M})$ in the conventional MC method to $O(1/M)$, where M is the number of samples (Caflisch 1998). This means that a much smaller number of samples is required to achieve a 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 $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_M\}$: for example, 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 probability density function (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; for example, 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, $\{z^m_{1:T}\}^M_{m=1}$. We create the samples $\{\zeta^\ell\}^L_{\ell=1}$ in a similar way.

Notably, even though our method still requires sampling from a high-dimensional space, the quasi-Monte Carlo 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 (the traditional sample average method) to achieve more accurate approximation in Equation (4) and speed up the solution. The feasibility of using the LHS method benefits greatly from the multivariate Gaussian distribution because it allows for an explicit form of the PDF and independent sampling for each dimension. This advantage would not be possible if we were 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 *reused* to avoid resampling from X when the forecast is updated, and the optimization is solved again each day (step 2 in the DSS). The mapping from Z to X is an exogenous input that can be trained offline (for example, on a 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. 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.

Implementation of Delta Coverage and Practical Challenges

In this section, we outline our tiered implementation process when deploying an analytics-based solution in a healthcare environment, which may also apply to other researchers in similar endeavors. We detail the challenges that we faced during the pilot in Appendix F.

The Delta Coverage analytics suite was launched in October 2021 as a Microsoft PowerBI application (details of this dashboard are in Appendix E). Because of the novelty of the program, we had no benchmark examples. To mitigate potential risks, we executed a three-phase tiered implementation with report outs to gain buy-in from upper-level management after each phase.

Preimplementation: Counterfactual

Before implementation, we conducted a counterfactual analysis using two 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 because it demystified the "black-box" DSS and showcased the power of OR analytics. This was especially valuable given the previous experiences of IUH 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 five months, we field tested the system live, running it daily to estimate the full-time equivalent staff 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 senior 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 that the

program is made operational such that it can 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 IUH's chief nursing officers. 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. We describe the reasons behind the delay between the prototype and launch along with other practical challenges in Appendix F.

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 because some of the data cannot be collected after the fact; for example, data that come from central data warehouses might be 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. See the details of the implemented dashboard in Appendix E.

Conclusion

The statewide Delta Coverage Program presents a collaborative effort between academia and industry, and it is an important first step in 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. Notably, the success of the pilot goes beyond addressing immediate staffing concerns; it demonstrates the feasibility of a new approach to nurse staffing. Historically, the healthcare industry has heavily relied on travel nursing to address staffing gaps, a practice fraught with logistical and financial challenges. The Delta Coverage pilot, despite its seemingly modest scale, serves as a proof of concept for a more sustainable and efficient solution. By integrating full-time resource nurses capable of providing care in multiple hospitals and adapting to short-notice staffing needs, this innovative approach shows that it is possible to reduce reliance on costly and inflexible travel nursing contracts. This 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 it serves.

Appendix A. Model Specifications and List of Notations

Table A.1. List of Notations for the Stochastic Optimization Model

Notation	Description
T	Length of the planning horizon for on-call and deployment decisions
k	Number of hospitals in the system
d_t^i	Number of DC nurses scheduled to work at time t who have home hospital i
$\mathbf{X}$	A $k \times T$ random vector denoting the demand for nurses at hospitals $i = 1, \ldots, k$ in time period $t = 1, \ldots, T$
a_t^{ij}	Decision variable denoting how many nurses to put on call for travel from location i to location j at time t
b_t^{ij}	Decision variable denoting how many nurses to deploy from location i to location j at time t (this variable depends on the realization of the nurse demand random vector, $\mathbf{X}$, as the decision is made after observing the census at all the hospitals)
n_t^i	The number of nurses initially scheduled at hospital i at time t prior to DC deployment
$\overline{n}_t^i$	The total (net) number of nurses at hospital i at time period t after deployment decisions, b_{ij}^t, have been executed
S^{ij}	The number of consecutive shifts a nurse must work at hospital j having been transferred from hospital i
c_u	Unit nurse understaffing cost
c_t^{ij}	Cost of putting a nurse on call for travel from hospital i to hospital j at time t
c_p	Cost of premium pay for nurses who are working at a remote hospital
η	The amount of additional travel cost recouped by canceling a deployment

A.1. Stochastic Optimization Model for DC Program Decision Support

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 activating the transfer of an on-call nurse or canceling 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*.

Nursing shortage is captured via the understaffing cost, c_u. The cost associated with the transfer decision $\mathbf{a}$ has two components: (1) the fixed cost that compensates for the transfer c_t^{ij}, which depends on the transfer distance, and (2) the variable cost that compensates for the length of the secondment $c_p S^{ij}$. If a transfer is cancelled during recourse, we recoup $1-\eta$ percentage of the transfer cost. Mathematically, the objective is

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

$$V(\mathbf{a},\mathbf{b},\mathbf{X}) = \min_{\mathbf{b}} \sum_{t=1}^{T}\sum_{i=1}^{k}\sum_{j=1}^{k}[c_u(X_t^i-\overline{n}_t^i)^+ - (1-\eta)(c_t^{ij}+c_p S^{ij})(a_t^{ij}-b_t^{ij})^+], \tag{A.2}$$

subject to

$$\sum_{j=1}^{k} a_t^{ij} \le d_t^i - \sum_{j=1}^{k}\sum_{\ell=(t-S^{ij}+1,1)^+}^{t-1} a_\ell^{ij}, \quad \sum_{j=1}^{k} b_t^{ij} \le \sum_{j=1}^{k} a_t^{ij}, \qquad \forall i,t, \tag{A.3}$$

where d_t^i is the number of available DC nurses with home location i on day t and $\overline{n}_t^i$ 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:

$$\overline{n}_t^i = n_t^i - \sum_{j=1}^{k}\sum_{\ell=t-S^{ij}}^{t} b_\ell^{ij} + \sum_{j=1}^{k}\sum_{\ell=t-S^{ji}}^{t} b_\ell^{ji}. \tag{A.4}$$

Appendix B. More Details on the Generative Model

B.1. SDE for Modeling Generative Dependence Structure

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

$$A_0 = a_0; \quad D_0 = d_0;$$
$$A_t = A_{t-1} + b_a(A_{t-1}) + \sigma_a Z_t^a, t = 1,\dots,T \tag{B.1}$$

$$D_t = D_{t-1} + b_d(D_{t-1}) + \sigma_d Z_t^d, t = 1,\dots,T, \tag{B.2}$$

where the sequences of latent variables $Z_1^a,\dots,Z_T^a \sim^{iid} \mathcal{N}(0,I_k)$ and $Z_1^d,\dots,Z_T^d \sim^{iid} \mathcal{N}(0,I_k)$ are all independent and identically distributed standard Gaussian vectors in $\mathbb{R}^k$ and drive the arrival and discharge processes. Equations (B.1) and (B.2) 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.

B.2. VAE Learning Framework

Instead of directly evaluating the likelihood function $p_\theta(\gamma_{1:T})$ given in Equation (3), VAE optimizes the ELBO as the surrogate

Table A.2. List of Notations for the Prediction Model

Notation	Description
X_t	A vector that corresponds to the patient census (number of patients) on day t in k hospital units
A_t	Daily arrivals to the hospital
D_t	Daily discharges from the hospital
ϵ_t	The measurement errors
Δ_t	The difference between arrival and discharge variables A_t and D_t (for example, the net changes in X_t's); that is, $\Delta_t = A_t - D_t$
Z_t^a	Sequence of latent random variables driving the arrival process, A_t
Z_t^d	Sequence of latent random variables driving the departure process, D_t
Z_t	Sequence of latent variables driving the cumulative differences between arrivals and departures
$z_{1:T}$	Sequence of *realized* latent (prior) variables
Γ_t	The cumulative difference between the census on day t and the initial census $X_0 = x_0$
$\gamma_{1:T}$	The observed cumulative difference sequence from data
θ	The set of parameters to be learned to forecast the census
$p_\theta(\gamma_t, z_{1:t})$	The conditional distribution of the cumulative difference as a function of the latent variables z
$q_\phi(z_{1:T}\|\gamma_{1:T})$	The variational distribution that approximates the true posterior distribution
$f_\theta(\cdot)$	The parameterized decoder function with parameter θ to learn conditional distribution p_θ
$f_\phi(\cdot)$	The parameterized encoder function with parameter ϕ for the variational distribution q_ϕ

objective derived in our setting:

$$\begin{aligned}
\log p_\theta(\gamma_{1:T}) &= \log \int p_\theta(\gamma_{1:T}, z_{1:T}) dz_{1:T} \\
&= \log \int p_\theta(\gamma_{1:T}, z_{1:T}) \frac{q_\phi(z_{1:T} \mid \gamma_{1:T})}{q_\phi(z_{1:T} \mid \gamma_{1:T})} dz_{1:T} \\
&\geq \mathbb{E}_{z_{1:T}\sim q_\phi}\left[\log\left(\frac{p_\theta(\gamma_{1:T}, z_{1:T})}{q_\phi(z_{1:T}|\gamma_{1:T})}\right)\right] \\
&= \mathbb{E}_{z_{1:T}\sim q_\phi}\left[\log\left(\frac{\prod_{t=1}^T p(\gamma_t|z_{1:t}))p(z_t|z_{1:t-1})}{\prod_{t=1}^T q_\phi(z_t|z_{1:t-1}, \gamma_{1:t})}\right)\right] \\
&= \sum_{t=1}^T \mathbb{E}_{z_{1:t}} \log p(\gamma_t|z_{1:t}) \\
&\quad - \mathbb{E}_{z_{1:t-1}} D_{KL}(q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}) \| N(0, I)) \\
&= \mathcal{L}(\gamma_{1:T}).
\end{aligned} \tag{B.3}$$

Recall that the key for VAE evaluation comprises 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 because 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 Kullback-Leibler 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 we 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 we will use z^{post} to denote samples from the posterior distribution.

B.2.1. Decoder. A key step in deriving the ELBO in Equation (B.3), particularly from line 3 to line 4, is via the following decomposition:

$$\begin{aligned}
p_\theta(\gamma_{1:T}, z_{1:T}) &= p_\theta(\gamma_{1:T}|z_{1:T}) p(z_{1:T}) \\
&= \left(\prod_{t=1}^T p_\theta(\gamma_t|z_{1:t})\right) p(z_{1:T}) \\
&= \prod_{t=1}^T p_\theta(\gamma_t|z_{1:t}) \prod_{t=1}^T p(z_t|z_{1:t-1}),
\end{aligned} \tag{B.4}$$

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 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 variables $(\gamma_w|z_{1:w}) \perp (\gamma_v|z_{1:v})$ 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 $p_\theta(\gamma_t|z_{1:t}) \sim N(\mu_{t,\theta}, \sigma_{t,\theta})$, 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 because the *difference* in census can be either positive or negative (in contrast to arrival and departure times, which must 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. Although the priors are assumed to be independent, the decoder f_θ allows us to capture the underlying complex correlations.

B.2.2. Encoder. We factor the variational distribution $q_\phi(z_{1:T}|\Gamma_{1:T})$ as

$$q_\phi(z_{1:T}|\gamma_{1:T}) = \prod_{t=1}^T q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}). \tag{B.5}$$

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 because we need to condition 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 $q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}) \sim N(\mu_{t,\phi}, \sigma_{t,\phi})$ is also 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.

B.2.3. Decoder Design. For the generative process, Difference-learning Timeseries Variational Autoencoder (DT-VAE) uses a decoder $f_\theta(\cdot)$ with parameter θ to decode latent variables $z_{1:t}$ to generate γ_t. That is, the decoder $f_\theta(\cdot)$ learns the conditional distribution $p_\theta(\gamma_t|z_{1:t})$. Recall that a key step in deriving the ELBO in Equation (B.3), particularly from line 3 to line 4, is via the decomposition for $p_\theta(\gamma_{1:T}, z_{1:T})$ as given in Equation (B.4), 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 (B.4), 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 $z_{1:t}$. This essentially makes γ_t *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 variables $(\gamma_w|z_{1:w}) \perp (\gamma_v|z_{1:v})$ 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 $\prod_{t=1}^T p_\theta(\gamma_t|z_{1:t})$.

For the prior distribution, we assume they are independent Gaussian, namely $p(z_t|z_{1:t}) \sim N(0, I)$. Although 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 the information $z_{1:t}^{prior}$ from all time steps recursively, with a feed-forward network f_{θ_2}, further transforming the input to $\mu_{t,\theta}$ and $\sigma_{t,\theta}$. We denote

$$h_{t,\theta_1} = f_{\theta_1}(h_{t-1,\theta_1}, z_t) \qquad (\mu_{t,\theta}, \sigma_{t,\theta}) = f_{\theta_2}(h_t), \tag{B.6}$$

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

B.2.4. Encoder Design. DT-VAE learns an encoder $f_\phi(\cdot)$ with parameter ϕ to encode observed $\gamma_{1:t}$ into the variational (posterior) distribution $q_\phi(z_{1:T}|\gamma_{1:T})$. Recall that the posterior distribution $q_\phi(z_{1:T}|\Gamma_{1:T})$ can be written as Equation (B.5). 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 $q_\phi(z_t|z_{1:t-1}, \gamma_{1:t}) \sim N(\mu_{t,\phi}, \sigma_{t,\phi})$, a Gaussian distribution with a diagonal covariance matrix, where $\mu_{t,\phi}$ and $\sigma_{t,\phi}$ are learned using the encoder f_ϕ. To capture the reliance of historical information on both z^{post}'s and γ's, we decompose f_ϕ into three functions with parameters ϕ_1, ϕ_2, and ϕ_3:

$$\begin{aligned} h_{t,\phi_1} &= f_{\phi_1}(h_{t-1,\phi_1}, \gamma_t) \\ \mu_{t,\phi} &= f_{\phi_2}(h_{t,\phi_1}, \mu_{t-1,\phi}) \\ \sigma_{t,\phi} &= f_{\phi_3}(h_{t,\phi_1}, \sigma_{t-1,\phi}), \end{aligned} \tag{B.7}$$

where h_{t,ϕ_1} is the hidden state in RNN structure f_{ϕ_1}. For each time step, h_{t,ϕ_1} encodes 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 resemble the conditional structure in $q_\phi(z_t|z_{1:t-1}, \gamma_{1:t})$ from Equation (B.5). Similarly, the RNN structure f_{ϕ_3} outputs $\sigma_{t,q}$ by utilizing h_{t,ϕ_1} and $\sigma_{t-1,\phi}$, which contain prior information of $\gamma_{1:t}$ and $z_{1:t-1}^{post}$. 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.

B.2.5. Computational Time. The DT-VAE method typically requires 500–1,500 epochs in training. This is far fewer than the 5,000–10,000 training epochs required by TimeGAN. The actual training time of DT-VAE varies by the training data size. Data sets with around 500 sample paths require about 10–20 minutes of training and about 5 minutes to generate 1,000 sample paths. This is significantly more efficient compared with TimeGAN, which can take at least three hours for training.

Appendix C. Prediction Performance Evaluation

We demonstrate the advantage of our method (the generative modeling structure and cumulative-difference learning) over traditional statistical methods, such as Autoregressive (AR) models. Our evaluation platform is a semisynthetic hospital census data set 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 process (Cox et al. 2005), with the drift function depending on the day of week and the daily discharges $d(t)$ coming from simulating patient movements within hospital units. All the parameters to simulate the arrivals and discharges are calculated empirically using real data. We provide an overview of the CIR model, a description of the real data set, and details of the semisynthetic generation in the rest of this section.

C.1. Cox–Ingersoll–Ross Model

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

$$dr(t) = \alpha(\mu - r(t))dt + \sigma\sqrt{r(t)}dW(t), \tag{C.1}$$

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 show that the hospital arrivals exhibit a strong day-of-week pattern. We describe how we modify the standard CIR process to generate a time-varying drift function in Appendix C.2.

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

$$\begin{aligned} r(t) = \max(&r(t-1) + \alpha(\mu - r(t-1))\Delta t \\ &+ \sigma\sqrt{|r(t-1)|}\sqrt{\Delta t}z_t, 0), \end{aligned} \tag{C.2}$$

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

C.2. Description of the Real Data Set from IUH

The real data set comes from an IUH hospital in Indiana. The data set contains patient-level movement history between different units in the hospital. The data span 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 contain time stamps on arrival time to each unit, the transfer in/out times between units, and the discharge time from the hospital. Using 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 $a_{u,t}$ denotes 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, we denote

- *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*: $p_u = 1/T\sum_{t=1}^{T}\frac{a_{u,t}}{a_{los,t}}$ for each u; and
- *Length of Stay (LOS) distribution*: $p_{u,s}^{dis} = \frac{X_{u,s}}{X_u}$;

where $X_{u,s}$ denotes the number of patients staying in unit category u for s days and X_u denotes the total number of patients staying in this unit category. For the LOS distribution, we further assume that the maximum LOS is four days (validated by the data because the proportion of patients staying longer than four days is minimal). With these parameters estimated empirically from the real data set, we then use them to generate semisynthetic data in Algorithm C.1.

Algorithm C.1 (Semisynthetic 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:

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

Assign arrivals to units. For each of unit $u \in \{MS, ICU\}$,

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

Generate discharges. Using 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. Generating census according to

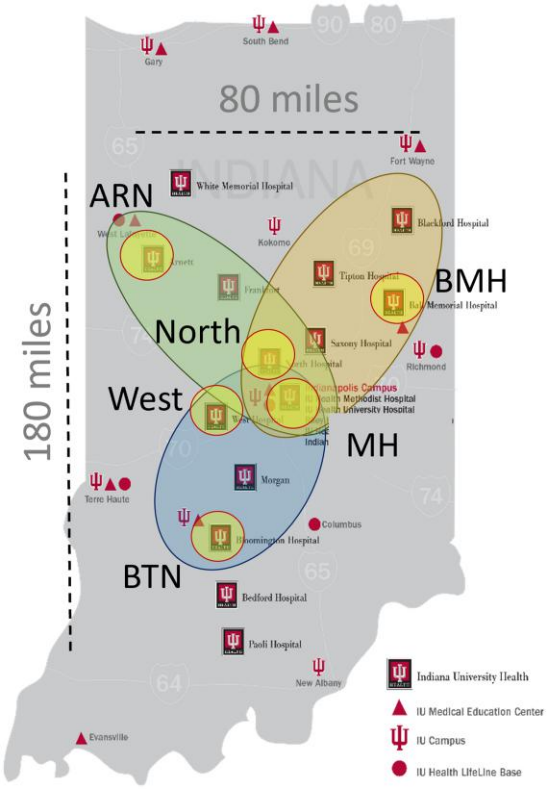
$$x_u(t) = x_u(t-1) + a_u(t) - d_u(t).$$

Algorithm C.1 describes the procedure of generating the semisynthetic data. Note that to capture the day-of-week pattern in the arrival rates, we modify the standard CIR process to generate a time-varying drift function, where μ_i follows a periodic pattern with one week (seven days) as the period. Correspondingly, we need to adjust the mean reversion factor α_t to be time varying through a weekly update scheme. For example, set $\alpha_1,\ldots,\alpha_7 = 0.1$ 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 semisynthetic generation via Algorithm C.1 is that it allows us to calculate the "ground truth" parameters, such as the expected daily number of arrivals and discharges. Using these calculated numbers, we could compare them with the corresponding results estimated from the generative models for evaluation.

Appendix D. Enlarged Figure for Delta Coverage Network Design

In this appendix, we discuss the final design of the Delta Coverage program as implemented at the Indiana University Health System. The small circles in Figure D.1 are participating hospitals. The larger circles Algorithm C.1 are the pods of hospitals, each with its own Delta Coverage team. A Delta Coverage team only floats within its own pod.

Figure D.1. (Color online) The Final Network Configuration for the Delta Coverage Program



Note. IU, Indiana University.

Appendix E. Delta Coverage Dashboard Functionality and Features

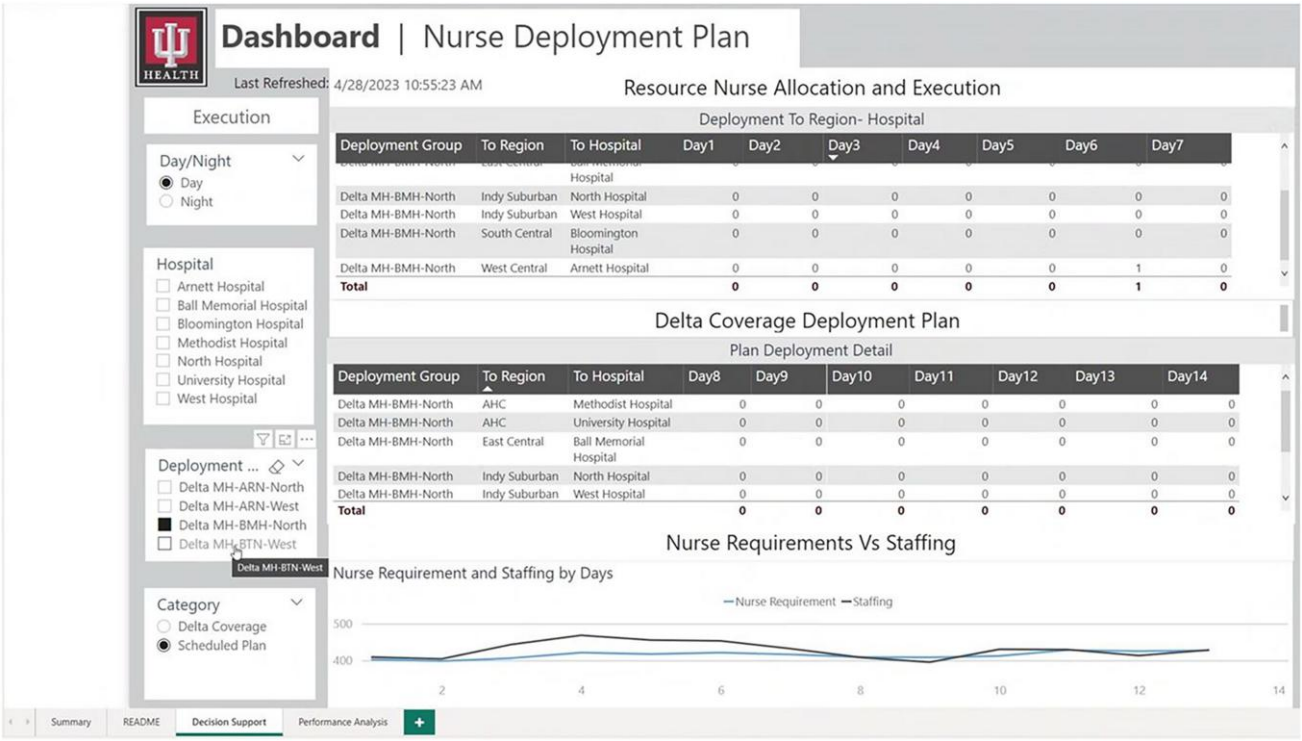
In this section, we describe the Delta Coverage Dashboard and how it supports on-call and deployment decision making in a variety of ways.

E.1. Dashboard Functionality and Usage

Once or twice a day, non-DC staffing data for the next 21 days are pulled from the Kronos timekeeping database and from a separate DC staffing database. The latter was manually curated to allow the DC program's implementation team to have more control over the data stream as the program was being rolled out. We also pull patient location data from the enterprise data warehouse; these data provide information about individual patient movement for the past 30 days. The movement data contain the location of each patient at 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 levels (medical/surgical, progressive care unit (PCU), or ICU) at each hospital. The data are gathered separately for day versus night shifts, with day-shift data starting from 11 a.m. and night-shift data starting from 11 p.m.

Once the data pass through the pipeline, they are entered into the prediction model, which can generate census sample

Figure E.1. (Color online) This Screenshot Shows the Full View of the Delta Coverage Dashboard Front Page



Source. IU Health Delta Coverage PowerBI Dashboard, Jacob Cecil.

paths to input into the stochastic optimization model. The optimization run uses a warm-start approach. That is, the algorithm starts from 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 entered into a platform-agnostic comma separated value (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, as shown in Figure E.1. The interface allows the user to display different views of the data in graphical form. For example, the lower right panel of Figure E.1 plots the nurse demand versus the staffing. It also allows the user to filter the table based on the selected criteria. The user can select day or night shift (in the upper left panel of Figure E.1), any subset of hospitals (in the second panel down the left side of Figure E.1), and the deployment group, which denotes the set of DC nurses being considered for transfer.

The DC nurse manager deploys DC nurses scheduled for the current day based on the optimization model's deployment suggestions. Once a week, the DC manager informs the DC nurses of their planned work (on-call) locations based on the optimal on-call decisions generated by the most recent run of the full optimization model.

E.2. Dashboard Visualization Features

One of the key features of the Delta Coverage Analytics application is a suite of visualizations to help users understand the impact of the nurse deployment actions on the broader system. The visualizations allow for

1. heat maps detailing the level of understaffing at all the hospitals before and after the deployment decisions and
2. 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 because they allow

1. users to test what-if scenarios and get immediate feedback on how changing the optimal recommendations would impact the system and
2. management to provide evidence to the individual hospitals of why the decisions are being made and how the decisions increase fairness in the system.

As an example, the lower right panel of Figure E.1 displays the forecasted demand and scheduled nurses over the next two weeks. In the lower left panel of Figure E.1, users can adjust which staffing plan to view. On the main page (not displayed here), they can view the staffing and demand based on the current schedule or the recommended schedule after the optimal deployment decisions. Therefore, managers can immediately see the impact of the optimization recommendation.

Appendix F. Detailed Postpilot Analysis and Lessons Learned

Our system included three phases of performance runs that are associated with the three phases of implementation. In the preimplementation phase (historical counterfactual), the two-month analysis suggested a 4% reduction in understaffing. In phases 1 and 2 (the live testing and tuning of the analytics

suite), we projected that the Delta Coverage program could potentially reduce understaffing by 5% and reduce 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. To perform the analysis, we compared two cases for a fair "apples to apples" assessment of the pilot program.

Case F.1. We counterfactually assigned each Delta Coverage nurse to a fixed hospital location ("home hospital") and did not allow that nurse to work at any other hospital (i.e., standard pre-Delta Coverage approach).

Case F.2. We compare the counterfactual results with the actual implementation results from the pilot (which involved moving nurses based on the Delta Coverage analytics tool).

From the actual historical data, we were able to pull the staffing schedule, including the number of unit-based nurses, resource nurses, and travel nurses who worked and the number of patients by acuity (ICU, PCU, and medical/surgical). We calculated the number of nurses needed on each shift by taking the number of patients of each acuity and dividing by the industry standard patient-to-nurse ratio for that acuity level. For example, if there are 10 ICU patients (two to one ratio for ICU), 24 PCU patients (three to one ratio for PCU), and 100 medical/surgical patients (five to one ratio for medical/surgical), the number of nurses required would be $10/2 + 24/3 + 100/5 = 33$. We calculate the patient census at 11 a.m. for the day shift and 11 p.m. for the night shift. To calculate the amount of understaffing, we subtract the number of nurses working in the hospital on a given shift from the number of nurses required in that hospital on that shift. In the previous example, if there were 32 nurses working, then the understaffing would be $33 - 32 = 1$ nurse. We truncate understaffing at zero so that if there had been 34 nurses in the example, then understaffing would be $(33 - 34)^+ = 0$. The amount of overstaffing is calculated similarly.

For each shift that a Delta Coverage nurse worked, we compared the actual amount of understaffing that occurred with the amount of understaffing that would have occurred had the Delta Coverage nurse worked the shift in that nurse's counterfactual "home hospital." We utilized the same method for overstaffing. In Appendix F.1, we present the impact of Delta Coverage on the system as a whole by calculating understaffing and overstaffing metrics across all Delta Coverage shifts in all participating hospitals.

F.1. System-Level Metrics

F.1.1. Phases 1 and 2. To test our system prior to implementation, we pulled the most recent two months of historical data with staffing and patient census for each hospital on each shift. We then ran the model iteratively starting with the first date in the data set. Specifically, we provided the model with the staffing and patient census for the current date (starting with the first date) and the future staffing schedules for the Delta Coverage planning horizon (e.g., the next three weeks). We then ran the model to determine the Delta Coverage nurse deployment decisions. We then added the counterfactual Delta Coverage nurses to the staffing plan for all the days and shifts covered by the Delta Coverage planning horizon. We incremented the date by one, moving to the next day in the data. Using the Delta Coverage deployment decisions, we then calculated the understaffing and overstaffing for this subsequent day as if the Delta Coverage tools' plan had, in fact, been implemented. We continued running the tool on each consecutive day until reaching the end of the data and then summed the understaffing and overstaffing over all shifts in the historical data. For comparison, we created a second counterfactual in which the Delta Coverage nurses were instead assigned to a single hospital and not allowed to deploy to other hospitals. Using the same number of (counterfactual) nurses on each shift along with their hospital assignments, we modified the staffing plan by adding those nurses to the shifts at their assigned home hospital and calculated the understaffing and overstaffing on each shift. We then added the measures over the entire time horizon. In this experiment, we assigned home hospitals to the non-DC counterfactual nurses by spreading them evenly across the six pilot hospitals, with the larger hospitals assigned an additional nurse because the number of nurses was not evenly divisible by six.

F.1.2. Phase 3. The results of the pilot from May 7 to June 23, 2023 were better than our initial dry run had projected. In this analysis, we consider the impact that the Delta Coverage program has had on understaffing and overstaffing in terms of the number of understaffed shifts eliminated, the percentage reduction in understaffing, and the estimated annual cost savings from the program.

F.1.3. Understaffing. Among the shifts that the DC nurses worked, in a little more than one month (36 days), the Delta Coverage pilot reduced understaffing by 33.5 shifts, which is equivalent to

- a 17% reduction in understaffing and
- 340 fewer understaffed shifts per year (34 shifts per DC nurse per year).

We obtain the annual estimate by extrapolating from the 36-day pilot by estimating the daily reduction in understaffing and then multiplying by 365: that is, $365 \times 33.5/36 = 340$ annualized shifts. We acknowledge the limitations of this method given the potential for changes in system characteristics over the course of an entire year.

Next, we compare the efficacy of hiring DC nurses versus hiring travel nurses, which are traditionally used to cover supply-demand mismatches. This allows us to demonstrate the marginal impact of the Delta Coverage Analytics Suite by comparing the program's actual performance with the counterfactual performance of hiring 10 travel nurses instead of the 10 DC nurses. Although we use travel nurses as our example because they are typically hired to cover demand and staffing mismatches, the following analysis applies to hiring any type of non-DC nurse.

To execute our counterfactual, we use the staffing data for all of the days/shifts (day versus night) that the Delta Coverage nurse worked as well as data on the number of patients in the participating hospitals. We then create a simulated scenario in which Delta Coverage nurses work all of their shifts in their home hospitals instead of where they actually worked. Recall that travel nurses (non-DC nurses) do not move between hospitals, so the scenario described is

the equivalent of hiring 10 travel nurses into the DC nurses' home hospitals instead of the DC nurses who were actually utilized. For example, for a DC nurse whose home hospital is IUH Bloomington Hospital, every time that the nurse is scheduled for a shift, we increase the staffing level at Bloomington by one and reduce the staffing number where the nurse actually worked the shift by one. This simulates nurses working all their shifts at their home hospitals instead of traveling between multiple hospitals.

From the adjusted staffing schedules, we obtain the amount of understaffing and overstaffing that would have occurred if 10 travel nurses had been hired instead of the 10 DC nurses using the same method as in phases 1 and 2. Comparing the understaffing metrics, 10 travel (non-DC) nurses would have only reduced understaffing by nine shifts (in 36 days), which is equivalent to

- a 4% reduction in understaffing and
- 90 fewer understaffed shifts per year, equating to 9 shifts per non-DC nurse per year.

This is a much lower magnitude compared with the reduction of 33.5 shifts (17%) from the DC pilot, which translates into 340 fewer understaffed shifts annually. This projection demonstrates the substantial marginal benefit of hiring Delta Coverage nurses as opposed to travel nurses; hiring Delta Coverage nurses would result in $250 = 340 - 90$ fewer understaffed shifts than hiring traditional travel nurses. Stated differently, for every understaffed shift avoided by hiring a travel nurse, $340/90 = 3.7$ understaffed shifts would be avoided by hiring a DC nurse instead.

F.1.4. Staffing Cost. Next, we consider the financial implications of the Delta Coverage program. Specifically, we estimate the number of travel nurses who would need to be hired to achieve the same reduction in understaffing achieved by the 10 DC nurses over the course of the pilot. We begin by calculating the actual level and the travel nurse counterfactual level of understaffing at each hospital on each shift (day/night) for each day of the pilot. We then subtract the Delta Coverage understaffing from the travel nurse understaffing shift by shift. Thus, if understaffing at a given hospital on a given shift was better (lower) using the travel nurse staffing plan, the result is negative; conversely, if the Delta Coverage staffing plan was better, the result is positive. We then sum up the differences in understaffing at the hospital-shift (day/night) level across all days of the pilot to obtain the total differential in understaffing for each hospital. We calculated the total difference in understaffing as 24.5.

To determine how many shifts of understaffing are eliminated by each subsequent travel nurse addition, we take the conservative approach of assuming that new nurses will be assigned to all the currently understaffed shifts at their assigned hospitals. For example, if Methodist Hospital's night shift was understaffed by one shift on 5/21, two shifts on 5/29, and three shifts on 6/14 and if a new travel nurse was assigned to the Methodist night shift, then the total understaffing for the pilot would be reduced by three shifts, resulting in understaffing of zero shifts on 5/21, one shift on 5/29, and two shifts on 6/14. We continue adding nurses to the pilot hospitals until the total amount of understaffing is the same as the total understaffing during the Delta Coverage pilot. We add nurses to hospitals in two ways as described in the following paragraphs. We then count the number of nurses who were added to hospitals counterfactually to obtain the estimate of the number of nurses required to achieve the same understaffing as the 10 DC nurses.

F.1.4.1. "Crystal Ball" (Very Conservative). In this ideal situation, we assume that IUH has precise foreknowledge of the days, hospitals, and shifts that will experience understaffing. We then assign each subsequent non-DC nurse to the hospital, and we shift to achieve the maximum reduction in understaffing over the course of the pilot: that is, the hospital/shift-type combination (day versus night shift) that has the most days of understaffing given the current staffing situation. Once a nurse is assigned to a hospital/shift type, we reduce the understaffing on each understaffed day by one to simulate the nurse working all of the understaffed shifts in that hospital/shift type. After reducing the understaffing, we then find the next hospital/shift-type combination that has the most understaffed days, and we continue to add nurses until the total understaffing during the pilot is the same as that of the actual DC nurse pilot. Under this assumption, we retrospectively calculate the number of non-DC nurses required to eliminate these understaffed shifts. The result indicates that even if we were able to foresee the future, 16 non-DC nurses would be needed to achieve the same level of understaffing as the 10 DC nurses in our pilot.

F.1.4.2. More Realistic (Slightly Conservative). In this more realistic hospital/shift-type assignment method, to calculate the number of additional travel nurses required to achieve the same level of understaffing, we use the following procedure. Given that there is a desire to balance new hires across the main hospitals and shifts (day/night), we add new travel nurses to hospitals and shift types in an order that maintains a balance between the number of additional nurses assigned to each hospital/shift type. For example, suppose Methodist Hospital currently has no additional (counterfactual) night-shift nurses currently assigned, whereas all other hospitals have at least one. To balance the number of additional nurses assigned to each hospital, the next counterfactual nurse will be assigned to the Methodist night shift. Ties are broken randomly. The result of this analysis demonstrates that IUH would have to have hired 19 additional non-DC nurses to achieve the same level of understaffing that was achieved by the 10 DC nurses in our pilot. In terms of productivity, this implies that a Delta Coverage nurse is the equivalent of 1.9 travel nurses and also has the benefit of being familiar with the hospitals and care teams.

F.1.5. Overstaffing. On the other side of the staffing mismatch, consider overstaffing. Although hiring additional nurses can never decrease overstaffing, we show that our Delta Coverage program significantly mitigates the *increase* in overstaffing from additional hires. Consider again the scenario where travel nurses were hired instead of Delta Coverage nurses. First, note that incidents of overstaffing when travel nurses are on shift are particularly undesirable. Travel nurses cannot be low censused (i.e., the nurse is sent home if not needed) and must be paid for a full shift regardless of need. This results in excessive and unnecessary costs given

the high salaries that travel nurses command, and it often results in full-time nurses being sent home instead. This highlights another major benefit of the Delta Coverage program; comparing the overstaffing associated with hiring travel nurses versus Delta Coverage nurses, we find that the DC program has significantly lower overstaffing during the course of the pilot. Specifically, there were 29 fewer shifts in which overstaffing occurred during the pilot compared with the number of overstaffed shifts that would have occurred if travel nurses had been hired instead of Delta Coverage nurses, which projects to 289 fewer overstaffed shifts a year and a 43% smaller increase in overstaffing relative to having hired non-DC nurses instead.

F.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) that each nurse spends at a remote facility. Of interest is that (1) each Delta Coverage nurse has a sufficient variety of working locations; this is based on the feedback from these nurses that one of the reasons they joined the program is that they want to travel, but they also want the stability of working in their home hospitals. Additionally, (2) Delta Coverage nurses should have a similar amount of variety in their working locations to ensure that the travel regime is fair to all these nurses.

Figure F.1 provides a high-level visual summary of Delta Coverage nurses' work schedules. For the 10 individual nurses participating the six-week pilot, Figure F.1 shows the percentage of shifts that each worked at various hospitals. Some nurses worked in a pod of three hospitals, and others worked in a pod of two hospitals.

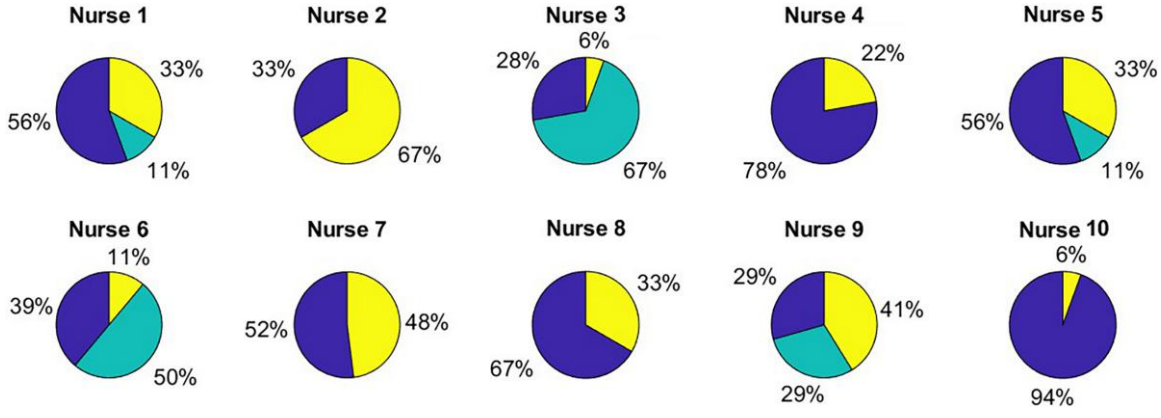
In general, we see a pattern that shows that the nurses have fairly similar distributions of work locations (we compare nurses in three-hospital pods separately from nurses in two-hospital pods). Recall that we do not need the shifts to be evenly distributed among hospitals but rather, that all nurses have a similar distribution of shifts across hospitals. As a final note, Nurse 10 was certified in one of three acuity levels, which somewhat restricted that nurse's transfer capability.

Additionally, we capture (1) the variety of opportunity (whether the nurses worked at each hospital often enough to earn a travel premium), (2) the stability of each nurse's schedule from week to week, and (3) the equity among Delta Coverage nurses for measures (1) and (2). We summarize these metrics in Table 1. We now explain more details about the calculation of these metrics. To measure 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 representing perfect inequality. In our context, a Gini coefficient of zero in terms of work variety means that the nurses spend an equal amount of time at each hospital in their catchment area. Similarly, if the nurse spends time in only one hospital, 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 because 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 that a Gini coefficient of 0.2–0.3 is considered very fair. With respect to equity between nurses, a smaller Gini coefficient means that an individual nurse's work variety and schedule stability are close to each other, indicating a fair implementation of the program. We use this interpretation of the Gini coefficient to evaluate our metrics as well. To measure stability, we calculate the coefficient of variation (CV) of each nurse's work variety from week to week. Specifically, we calculate the number of different hospitals at which each nurse worked in a week. We then calculate the mean and standard deviation of the weekly number of different hospitals worked across the five weeks and divide the standard deviation by the mean to obtain the CV.

F.2.1. 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 to June). The average work variety (mean of the Gini coefficient) across all Delta Coverage nurses is 0.42. Note that Nurse 10 was certified in only one acuity and thus, could not fill all nursing

Figure F.1. (Color online) The Pie Charts Show the Fraction of Shifts Worked at Each Nurse Location for the 10 Delta Coverage Nurses



roles. Thus, we remove this nurse when calculating the 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.

F.2.2. 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. Next, for each nurse, we calculate the CV of that nurse's work variety over the six-week horizon in 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 is. We adopt the convention that $CV < 1$ is considered to be low variability and that $CV > 1$ is considered to be high variability. Considering all nurses, the average CV of work variety is 0.41, and the equity (the Gini coefficient of the CV) is 0.31, indicating that the program is creating schedules that are stable, consistent, and fair across Delta Coverage nurses.

F.2.3. Delta Coverage Hospital Equity. To measure the fairness of the allocation of Delta Coverage nurses to hospitals, we again use the Gini coefficient. After removing the single outlier hospital (Bloomington Hospital (BTN)), which maintains the concept of fairness because BTN was well staffed during the pilot period, the Gini coefficient was 0.29, indicating a very fair allocation.

In summary, the previous analyses have demonstrated that the pilot not only achieved significant reductions in understaffing and overstaffing but also created nurse schedules and allocated Delta Coverage resources in a desirable and equitable manner.

F.3. Practical Challenges Encountered in the Pilot and Lessons Learned

F.3.1. 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 dashboard, which we completed testing in April 2022. However, the pilot launch was delayed until May 2023 because of 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.

Although 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 us 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 its 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: (1) relaunching and retesting the analytics suite and (2) recruiting nurses for the Delta Coverage program.

F.3.2. 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 not yet 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 are primarily designed for billing and finance. This required us to implement major work-arounds to ensure accurate operational conclusions. For example, we had to use patient location data (the location at which the patient is billed) to construct hospital occupancy data. However, we discovered a double-counting issue; numerous patients were mistakenly counted in two places because the inpatient beds were being held for them while they were in surgery or recovery rooms. Our team addressed these challenges through advanced planning, anticipating future transfers in the hospital, and incorporating an automated change detection mechanism.

F.3.3. Recruitment. As we mentioned, one of the major challenges and milestones was recruiting nurses for this novel program. 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 hyperparameters to include additional desirable features that various nursing teams mentioned in 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 involved running information sessions for DC-eligible nurses. Other design changes tested in the analytics suite included 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 the length of travel secondments (the number of 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 that the fraction of DC shifts allocated to each participating hospital was fair. The second wave of recruitment proved to be a success thanks to the implementation of design changes tested in the analytics suite.

References

Aiken LH, Sloane DM, Bruyneel L, Van den Heede K, Griffiths P, Busse R, Diomidous M, et al. (2014) Nurse staffing and education and hospital mortality in nine European countries: A retrospective observational study. *Lancet* 383(9931):1824–1830.

Allen LJS (2008) *An Introduction to Stochastic Epidemic Models* (Springer, Berlin), 81–130.

Allen LJ (2017) A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Model.* 2(2): v128–142.

American Hospital Association (2022) Massive growth in expenses and rising inflation fuel continued financial challenges for America's hospitals and health systems. Technical report, American Hospital Association, Washington, DC.

Anderson D, Bjarnadottir MV, Nenova Z (2022) Machine learning in healthcare: Operational and financial impact. Babich V, Birge JR, Hilary G, eds. *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management, vol. 11 (Springer, Cham, Switzerland), 153–174.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 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* 49(4):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 Environ. Res. Risk Assessment* 37(5):1839–1854.

Chan T, Park J, Pogacar F, Sarhangian V, Hellsten E, Razak F, Verma A (2021) Optimizing inter-hospital patient transfer decisions during a pandemic: A queueing network approach. Preprint, submitted December 21, http://dx.doi.org/10.2139/ssrn.3975839.

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

Desai A, Freeman C, Wang Z, Beaver I (2021) Timevae: A variational auto-encoder for multivariate time series generation. Preprint, submitted November 15, https://arxiv.org/abs/2111.08095.

Esteban C, Hyland SL, Rätsch G (2017) Real-valued (medical) time series generation with recurrent conditional GANs. Preprint, submitted June 8, https://arxiv.org/abs/1706.02633.

Flinkman M, Leino-Kilpi H, Salanterä S (2010) Nurses' intention to leave the profession: Integrative review. *J. Adv. Nursing* 66(7): 1422–1434.

Furukawa MF, Machta RM, Barrett KA, Jones DJ, Shortell SM, Scanlon DP, Lewis VA, O'Malley AJ, Meara ER, Rich EC (2020) Landscape of health systems in the united states. *Medical Care Res. Rev.* 77(4):357–366.

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

Griffiths P, Saville C, Ball J, Jones J, Pattison N, Monks T; Safer Nursing Care Study Group (2020) Performance of the Safer Nursing Care Tool to measure nurse staffing requirements in acute hospitals: A multicentre observational study. *BMJ Open* 10:e035828.

Hu Y, Chan CW, Dong J (2024) Prediction-driven surge planning with application in the emergency department. *Management Sci.*, ePub ahead of print May 24, https://doi.org/10.1287/mnsc.2021.02781.

Li T, Wu C, Shi P, Wang X (2024) Cumulative difference learning VAE for time-series with temporally correlated inflow-outflow. *Proc. AAAI Conf. Artificial Intelligence*, vol. 38, No. 12 (AAAI Press, Palo Alto, CA), 13619–13627.

Mogren O (2016) C-RNN-GAN: Continuous recurrent neural networks with adversarial training. Preprint, submitted November 29, https://arxiv.org/abs/1611.09904.

Owen AB (1998) Latin supercube sampling for very high-dimensional simulations. *ACM Trans. Model. Comput. Simulation* 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. Preprint, submitted November 6, https://arxiv.org/abs/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. *Internat. J. Nursing Stud.* 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 Oper. Management* 32(5):1433–1452.

Spetz J (2021) Leveraging big data to guide better nurse staffing strategies. *BMJ Quality Safety* 30(1):1–3.

Yoon J, Jarrett D, Van der Schaar M (2019) Time-series generative adversarial networks. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Proc. 33rd Internat. Conf. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY).

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

Jonathan E. Helm is a professor at Indiana University's Kelley School of Business. He is also the research co-director for the Center for the Business of Life Sciences. With experience at GE Healthcare and Mayo Clinic, he was a National Science Foundation Fellow for three years. His research has led to practical implementations, including a census forecasting system in Singapore, readmission reduction analytics in Indiana, and predictive analytics at Indiana University Health.

Pengyi Shi is an associate professor at the Mitchell E. Daniels, Jr. School of Business, Purdue University. Her research focuses on data-driven modeling and decision-making in healthcare and service operations. She has collaborated with major healthcare organizations in the United States, Singapore, and China. Her work has received multiple awards, including the IISE Outstanding Innovation in Service Systems Engineering Award in 2023 and the *MSOM* Responsible Research in OM Award in 2021.

Mary Drewes is the associate chief nurse executive at Indiana University Health, overseeing nursing operations for 16 hospitals and more than 9,000 nurses. She collaborates with senior executives and system leaders in information technology, finance, pharmacy, regulatory, quality and safety, and supply chain. She effectively manages cross-functional teams to address complex business challenges and serves on multiple boards to represent nurses.

Jacob Cecil is a senior data analyst at IU Health. He uses advanced analytical techniques to extract insights from complex datasets, enhancing healthcare delivery and patient experience. He has collaborated on developing a Python-based decision support model for staff placements. He also contributed to the Google Health Data Engine pilot project at IU Health, a cloud-based platform for real-time healthcare data analysis.