Appendix for "Operations (Management) Warp Speed: Rapid Deployment of Hospital-Focused Predictive/Prescriptive Analytics for the COVID-19 Pandemic"

Pengyi Shi, Jonathan E. Helm, Christopher Chen, Jeff Lim, Rodney P. Parker, Troy Tinsley, and Jacob Cecil

A Applications: Nurse Transshipment and Hiring

A.1 Literature on Transshipment

The literature on transshipment is extensive, and an excellent review appears in Paterson et al. (2011). The COVID related transshipment papers typically involve the movement of equipment, such as ventilators, using stochastic optimization (Bertsimas et al. 2021) and simulation (Bhavani et al. 2020) methodologies, or the movement of patients through diversion of new admissions (Parker et al. 2020) and transfers (Martucci et al. 2020). Our transshipment model differs in that we consider nurse transfers, which can be pre-planned and later either canceled or augmented with additional transfers after demand is realized. An interesting component that differentiates nurse transshipment is that nurses have a "home" location that they must return to when their distance-dependent secondment is complete.

A.2 Nurse Transshipment Model

We model nurse transshipment as a stochastic program, allowing both *ex-ante* ("planning") and *ex-post* ("emergency") transfers. The *ex-ante* decisions transfer a_t^{ij} nurses from location i to location j on day t of the planning horizon. After demand is realized, recourse actions occur in terms of b_t^{ijz}, which is the actual number of nurses transferred from i to j under scenario z. $b_t^{ijz} > a_t^{ij}$ represents an additional emergency transfer, whereas $b_t^{ijz} < a_t^{ij}$ represents a call-back to the nurses "home" location. Model notation and assumptions are listed below in Appendix A.3. The outer objective function is given by:

$$\min_{\mathbf{a}} \sum_{t=1}^{T} \sum_{i=1}^{L} \sum_{j=1}^{L} (pS^{ij} + t^{ij}) a_t^{ij} + \mathrm{E}\left[V(\mathbf{a}, \mathbf{b}, \mathbf{C})\right]$$

where the expectation is taken over the census profiles, $\mathbf{C}$, (scenarios indexed by $z = 1, ..., P$) and the recourse objective is:

$$V(\mathbf{a}, \mathbf{b}^z, \mathbf{C}^z) = \min_{\mathbf{b}^z} \sum_{t=1}^{T} \sum_{i=1}^{L} \left(\sum_{j=1}^{L} \left[(u_t p S^{ij} + t^{ij})(b_t^{ijz} - a_t^{ij})^+ - (1-\eta)(pS^{ij} + t^{ij})(a_t^{ij} - b_t^{ijz})^+ \right] + s_t^i (\Delta_t^i(\mathbf{b}^z, C_t^{iz}))^+ \right),$$

where $\Delta_t^i(\mathbf{b}^z, C_t^{iz})$ is the difference between demand (census workload) and supply (nurses):

$$\Delta_t^i(\mathbf{b}^z, C_t^{iz}) = C_t^{iz} - \left(x^i - \sum_{j=1}^{L} \sum_{k=(t-S^{ij}+1,1)^+}^{t} b_k^{ijz} + \sum_{j=1}^{L} \sum_{k=(t-S^{ji}+1,1)^+}^{t} b_k^{jiz} \right).$$

The primary constraints are capacity constraints: i.e., cannot transfer more nurses than are available at origin location ($i = 1, ..., L; t = 1, ..., T; z = 1, ..., P$):

$$\sum_{j=1}^{L} a_t^{ij} \leq x^i - \sum_{j=1}^{L} \sum_{k=(t-S^{ij}+1,1)^+}^{t-1} a_k^{ij},$$

$$\sum_{j=1}^{L} b_t^{ijz} \leq x^i - \sum_{j=1}^{L} \sum_{k=(t-S^{ij}+1,1)^+}^{t-1} b_k^{ijz}.$$

In addition to non-negativity constraints, auxiliary variables and constraints must be added to linearize the $(\cdot)^+$ operator, which allows the model to be solved efficiently via linear programming. In the outer objective, a nurse transferred from location i to location j will receive a salary premium p for the S^{ij} days she is seconded with an additional travel bonus t^{ij} based on distance. In the recourse objective, the salary premium is increased through the multiplier $u_t > 1$ for emergency transfers, whereas transshipment costs from the *ex-ante* plan are recovered for call-backs at the cost of percentage cancellation fee, η. Understaffing (= demand - nurses) at location i in period t incurs a unit cost s_t^i.

A.3 Transshipment Model Notation and Assumptions

The notation for the nurse transshipment model in Appendix A.2 is:

a_t^{ij} = the number of nurses transferred from location i to j at the start of period t, ex-ante

b_t^{ijz} = the number of nurses transferred from location i to j at the start of period t, ex-post, under census scenario z

C_t^{iz} = census (workload) at location i in period t under scenario z

x^i = number of nurses based at location i ("home")

L = number of locations

T = length of planning horizon

P = number of census scenarios

V = value of a recourse scenario

S^{ij} = length of secondment for a nurse sent from location i to j

Δ_t^i = difference between census workload and available nurses at location i in period t

p = daily payment premium for working away from "home"

t^{ij} = non-salary cost of transferring a nurse from location i to j

u_t = daily salary premium (multiplier) for emergency transshipment in day t

η = percentage fee for cancelling an ex-ante transfer

s_t^i = unit cost of shortage at location i on day t

The assumptions for the nurse transshipment model include:

- Only "home" nurses can be transferred. Any nurses already seconded are ineligible for additional transfers.
- Transfer decisions between locations for the T period planning horizon occur before the census profiles are known: a_t^{ij} for $t = 1, ..., T; i = 1, ..., L; j = 1, ..., L$.
- Each census profile consists of the workload (numbers of patients) at each of the L locations for each of the T periods. There are P scenarios.
- For each scenario z, transfer decisions between locations occur for the planning decision with the specific census profile known: b_t^{ijz} for $t = 1, ..., T; i = 1, ..., L; j = 1, ..., L; z = 1, ..., P$.
- The emergency adjustments upward (more nurses being sent in the recourse periods) will earn affected the nurses a premium of $u_t > 1$ in their premium pay. Thus, the cost for an upward emergency adjustment is $u_t p S^{ij} + t^{ij}$ with $u > 1$. This reflects the fact that nurses who are emergency transferred earn the premium for the entire duration of their secondment.
- The emergency adjustments downward (fewer nurses being sent in the recourse periods) will result in the earlier cost being recovered. Thus, the marginal cost for a downward emergency adjustment is $-(pS^{ij} + t^{ij})$ since those nurses will no longer be needed at the other location.

- A unit shortage cost of s_t^i is assessed for each unit difference between the census and available nurses at location i in period t.

A.4 Numerical Inputs for Nurse Transshipment Model

Table 1: Secondment days by region.

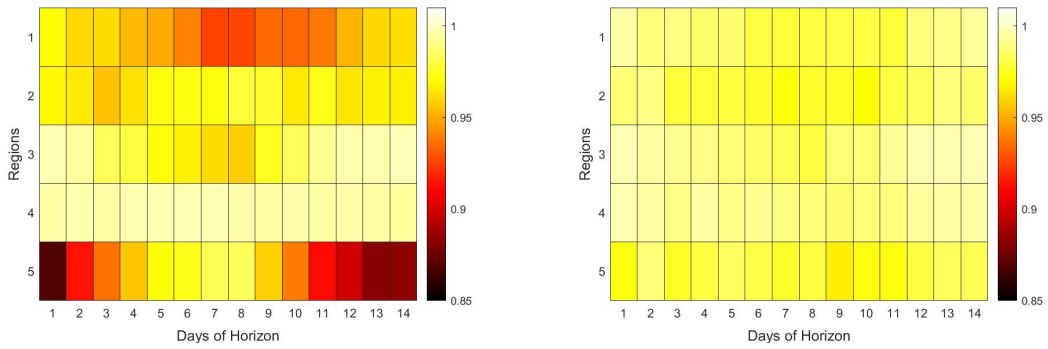
Region From/To	1	2	3	4	5
1	1	2	1	2	2
2	2	1	2	3	2
3	1	2	1	2	2
4	2	3	2	1	3
5	2	2	2	3	1

Table 2: Transfer cost matrix (proportional).

Region From/To	1	2	3	4	5
1	1	1.5	1.1	1.5	1.5
2	1.5	1	1.4	1.7	1.7
3	1.1	1.4	1	1.6	1.5
4	1.5	1.7	1.6	1	1.7
5	1.5	1.7	1.5	1.7	1

A.5 Further Operational Products of the Nurse Transshipment Model

In addition to the operational plan suggesting specific nurse transshipments between locations described in Section 5.1, another graphic that we produced helps IU Health better understand the disparities among the regions that occur as a natural consequence of expansion and semi-decentralized decision making. Figure 11b demonstrates the impact of the transfer program on region-wide understaffing by day compared with no transfers in Figure 11a. Darker colors indicate a greater level of understaffing.



(a) Without transfers (b) With transfers

Figure 11: Staffing to plan by region; percent of staffing needs met.

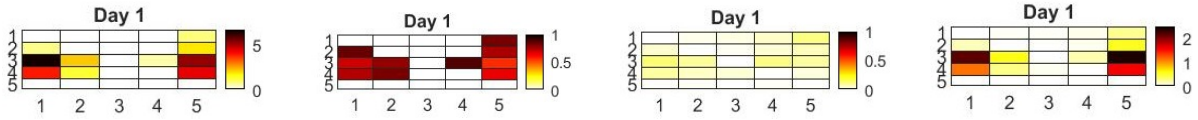
From the nursing management's perspective, these heatmaps can serve several purposes. They

provide a compact visual for potential future staffing shortfalls (as measured by percent under plan) across all regions simultaneously, which helps them make tactical and operational adjustments. They also provide data-driven justification to convince regions of the consistent fairness of the transshipment program compared with no transshipment. Finally, after initially being presented these results, there was significant interest in using this approach not only for transshipment but also for future hiring decisions to balance the system capacity (notice some regions are consistently short while others are more flush).

The final product that we delivered to IU Health is the day by day transshipment schedule, along with recourse actions to demonstrate the expected outcome of the initial plan. Figure 12 shows one day of the horizon where the y-axis is the "transshipment from" location and the x-axis is the "transshipment to" location. The entire 14 day plan is in Appendix A.6.

Analyzing these heatmaps provides some interesting insights into the strategy behind the transshipment plan. The initial plan, which is communicated to the nursing staff, proposes a significant volume of scheduled transshipments (Figure 12a) where around 50-80% are called back (in expectation) after demand is realized (Figure 12b) leading to very little need for emergency transshipment (Figure 12c), where the final plan results in transshipments that are on average a little less than a quarter of the planned transshipments.

One final interesting result is that the initial plan often exhibits a two-way phenomenon: on Day 1 there is a scheduled transshipment from Region 3 to 2 and from Region 2 to 3, simultaneously. At first this seems counterintuitive – why not take the net between the two and plan a one-way transshipment? We explained this phenomenon to IU Health in the following way. Initial transshipment is cheaper than emergency transshipment and inexpensive to reverse (both in the model and in reality). Thus, the initial plan pre-positions cheap backup capacity throughout the system to cover all of the probable scenarios. For example, if Region 3 has a spike in census, IU Health can pull back the planned transshipments *to* Region 2 (very cheap) and keep the pre-planned transshipments *from* Region 2 (cheaper than emergency transshipment) or vice versa if Region 2 has an unplanned spike. In a sense the initial plan creates phantom capacity with real benefit.



(a) Initial plan (b) % Call-backs (c) Emergency transfer (d) Final plan

Figure 12: Initial plan and expected recourse actions for unbalanced scenario.
y-axis: transshipment from location. *x-axis:* transshipment to location.

After sharing these results, IU Health confirmed that this is indeed an attractive, implementable approach to a transshipment program, largely avoiding last minute travel and limiting overall

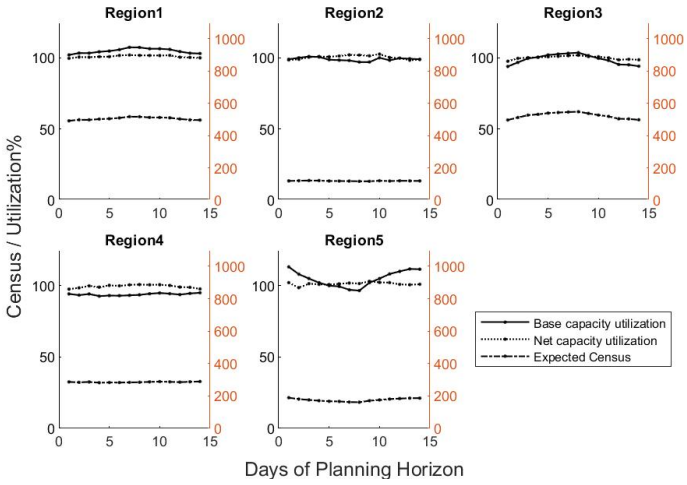


Figure 13: Expected capacity, census, and understaffing by day and region
Left y-axis: Capacity utilization. *Right y-axis:* Region census

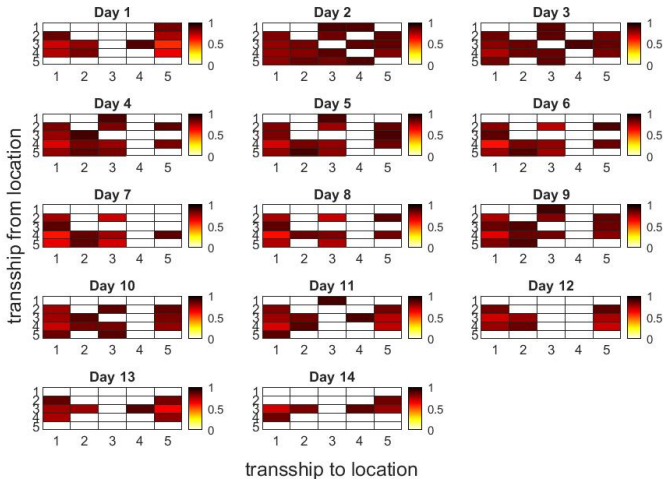


Figure 14: Expected percentage of call-backs.

system transshipments only to emergency need as a result of significant deviation from the forecast point estimate.

A.6 Complete Figures for Nurse Transshipment

In this section, we show the entire 14-day plan and corresponding performance metrics under the unbalanced setting in Figures 13-22.

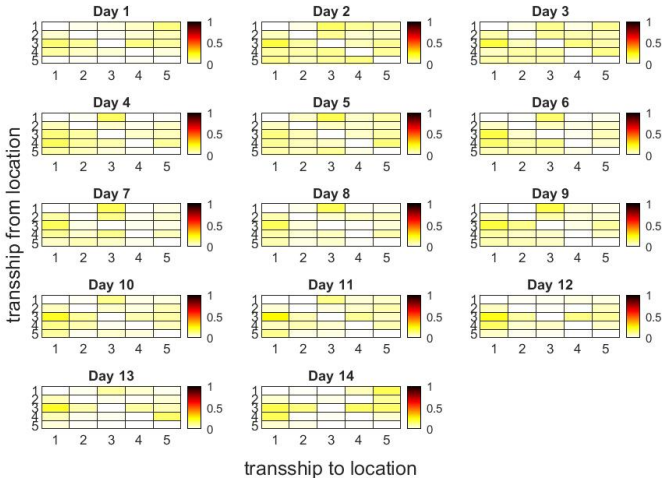


Figure 15: Emergency transshipment (ex-post).

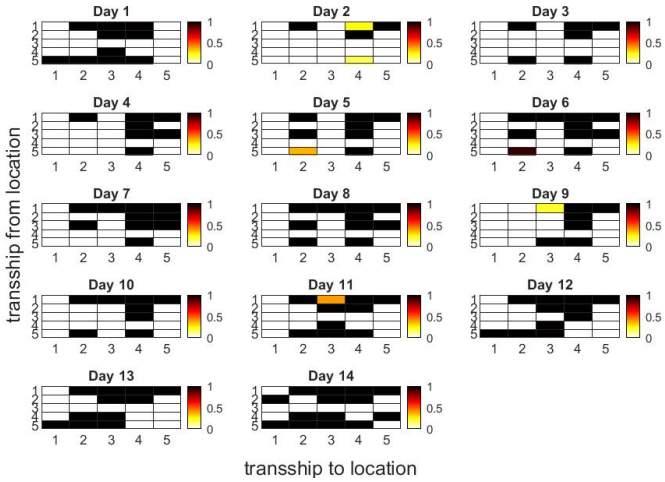


Figure 16: Percent emergency transshipments (percent above ex-ante).

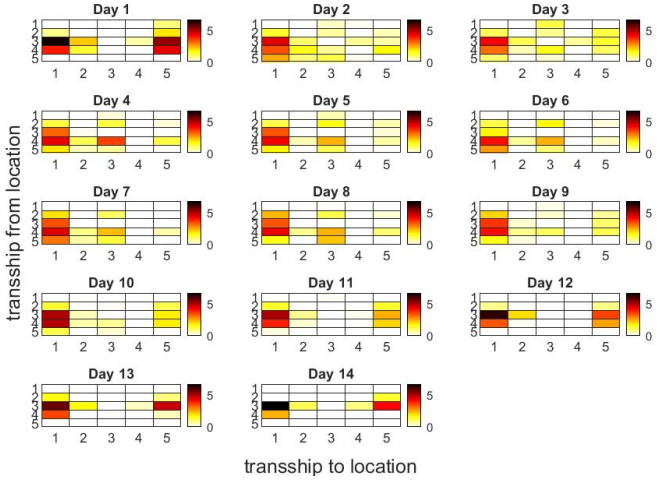


Figure 17: Initial transshipment plan (ex-ante).

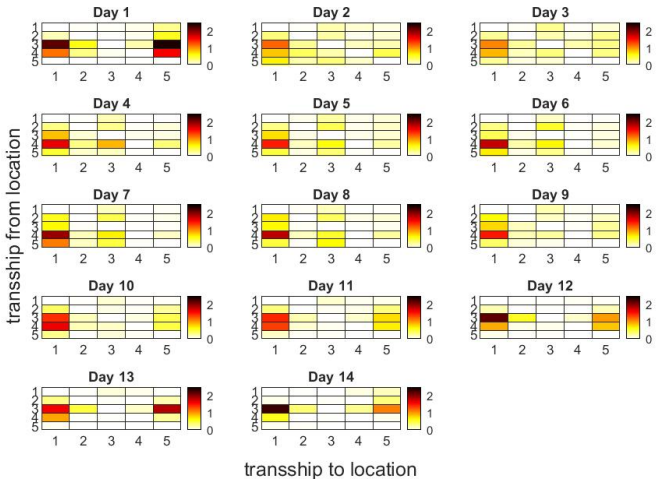


Figure 18: Expected final transshipments (ex-post).

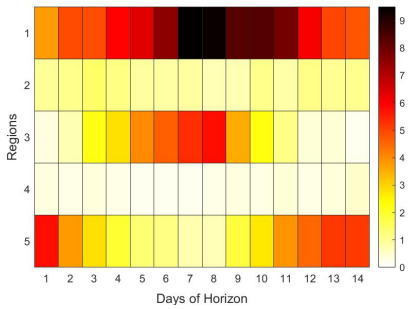


Figure 19: Expected understaffing by day and region with no transshipment.

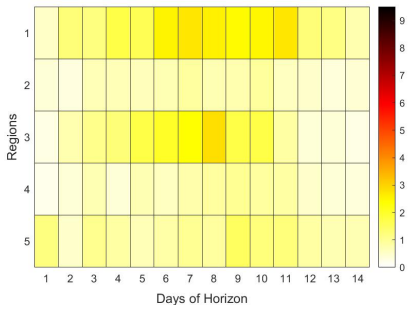


Figure 20: Expected understaffing by day and region with transshipment.

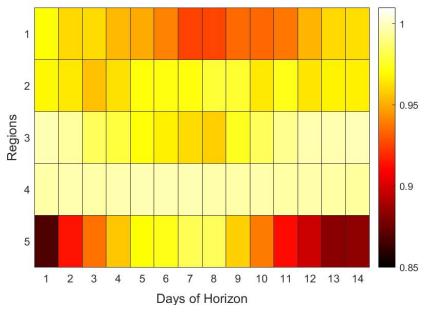


Figure 21: Expected percent staffing to need with no transshipping.

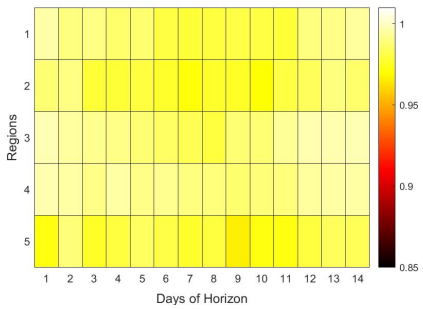


Figure 22: Expected percent staffing to need with transshipping.

A.7 Ventilator Transshipment Model

A ventilator transshipment model will resemble the nurse transshipment model in Appendix A.2 (we will retain the notation of Appendix A.3). All the ventilators are owned by the hospital network, so there is no urgency to return borrowed ventilators to their original location, unlike the borrowed nurses who are returned to their "home" location after a designated period. In this ventilator context, there is a regular transfer cost t^{ij} between locations i and j but no salary payment. In the recourse periods, the transfer cost is in period t is inflated by $u_t > 1$ to reflect the added expenses of the unplanned transfer. Likewise, if a previously planned transfer is reversed, a modest cancellation fee η is imposed.

The key difference with the nurse transshipment model is that ventilators can be transferred from any location in any period (if available), unlike the nursing application where a nurse already on secondment cannot be sent on another visit; this additional flexibility is reflected in constraints (16) and (17). Specifically, the total number of ventilators being sent (left hand side) cannot exceed the net number of ventilators available in that period (right hand side is original number of ventilators minus the total previously lent out plus the total previously borrowed). As before $a_t^{ij} \geq 0$ and $b_t^{ijz} \geq 0$ for $i, j = 1, ..., L$, $z = 1, ..., P$, and $t = 1, ..., T$,

$$\min_{\mathbf{a}} \sum_{t=1}^{T} \sum_{i=1}^{L} \sum_{j=1}^{L} t^{ij} a_t^{ij} + \mathrm{E}\left[V(\mathbf{a}, \mathbf{b}, \mathbf{C})\right],$$

where the expectation is taken over the census profiles, $\mathbf{C}$, (scenarios indexed by $z = 1, ..., P$) and the recourse objective is:

$$V(\mathbf{a}, \mathbf{b}^z, \mathbf{C}^z) = \min_{\mathbf{b}^z} \sum_{t=1}^{T} \sum_{i=1}^{L} \left(\sum_{j=1}^{L} \left[u_t t^{ij} (b_t^{ijz} - a_t^{ij})^+ - (1-\eta) t^{ij} (a_t^{ij} - b_t^{ijz})^+ \right] + s_t^i (\Delta_t^i(\mathbf{b}^z, C_t^{iz}))^+ \right),$$

with $\Delta_t^i(\mathbf{b}^z, C_t^{iz})$ being the difference between demand (census workload) and supply (ventilators):

$$\Delta_t^i(\mathbf{b}^z, C_t^{iz}) = C_t^{iz} - \left(x^i - \sum_{j=1}^{L} \sum_{k=1}^{t} b_k^{ijz} + \sum_{j=1}^{L} \sum_{k=1}^{t} b_k^{jiz} \right).$$

The constraints are capacity constraints; cannot transfer more ventilators than are available at each location ($i = 1, ..., L; t = 1, ..., T; z = 1, ..., P$):

$$\sum_{j=1}^{L} a_t^{ij} \leq x^i - \sum_{j=1}^{L} \sum_{k=1}^{t-1} a_k^{ij} + \sum_{j=1}^{L} \sum_{k=1}^{t-1} a_k^{ji}, \tag{16}$$

$$\sum_{j=1}^{L} b_t^{ijz} \leq x^i - \sum_{j=1}^{L} \sum_{k=1}^{t-1} b_k^{ijz} + \sum_{j=1}^{L} \sum_{k=1}^{t-1} b_k^{jiz}. \tag{17}$$

A.8 Nurse-hiring Newsvendor Model

Let r_j be the target patient-to-nurse ratio for hospital unit $j \in \{MS, ICU\}$. Using the offered-load approximation for the workload distribution from Section 3, we approximate the distribution of the workload in unit j on day t, $X_j(t)$, as a normal distribution $\mathcal{N}(x_j(t), v_j(t))$. Let q^* be the critical fractile (or service level). Through discussions with nursing management, we set $q^* = 0.67$, based on nurse overtime pay (double) that is typically incurred when understaffing occurs. This ratio is also consistent with previous literature (Green et al. 2013).

To determine the number of travel nurses to hire, we calculate the desired staffing level for unit j on day t, denoted by $N_j(t)$, based on the following equation

$$F_j(r_j N_j(t)) = \mathbb{P}(X_j(t) \leq r_j N_j(t)) = q^*,$$

with solution

$$N_j(t) = \frac{F^{-1}(q^*)}{r_j},$$

where F^{-1} is the inverse CDF of the workload distribution. From our offered-load approximation in Section 3.3, the marginal distribution of $X_j(t)$ follows the normal distribution, so the desired staffing level can be expressed as

$$N_j(t) = \frac{1}{r_j}\left(x_j(t) + z_{q^*}\sqrt{v_j(t)}\right),$$

where $x_j(t)/r_j$ is the average staffing level and $z_{q^*}\sqrt{v_j(t)}/r_j$ is the additional "buffer" against census variability.

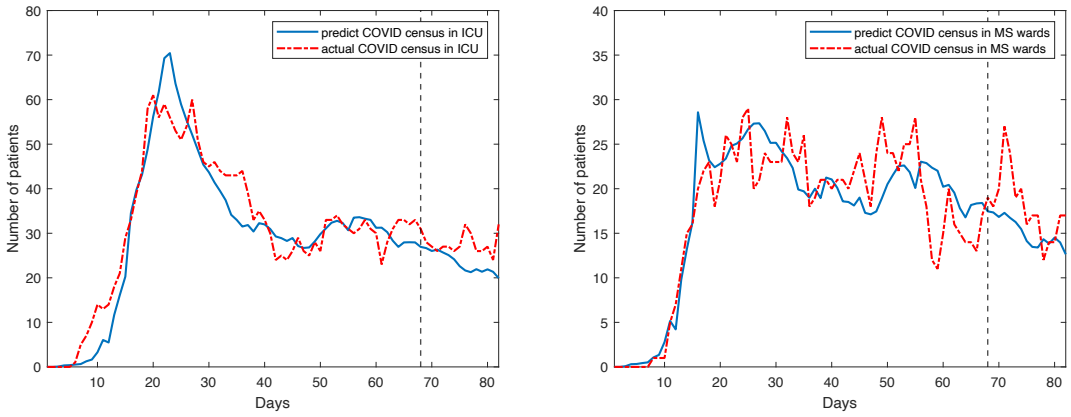
B Additional Development for Workload Prediction

B.1 Prediction Results from Other Months

Figure 23 shows the observed and predicted COVID inpatient census at one of IU Health's hospitals during the period of March 12, 2020 to June 2, 2020. The flow parameters are learned by using the data up to May 19, i.e., we use the observed census data from the first 68 days as the training data. Using the learned parameters, we generate a prediction over the next 14 days (May 20 to June 2), i.e., everything to the right of the vertical dashed line in Figure 23. The generated prediction is then compared with the observed census data to measure our prediction accuracy. In this period, the MAPE is 14.8% for ICU census, and 17.3% for MS census.

B.2 Offered-load Approximation for General Networks

We consider a general network with J stations. The workload on day $t+1$, denoted as the J-dimensional vector $X(t+1) = (X_1(t+1), X_2(t+1), \ldots, X_J(t+1))$, can be calculated recursively



(a) ICU census (b) M/S census

Figure 23: Prediction of patient census for the largest hospital in our partner healthcare network: March 12, 2020 - June 2, 2020. In the plots, census data from the first 68 days (left to the vertical dashed line) is used as the training data, and the remaining 14 days (right of the vertical dashed line) as the testing data.

as follows

$$X_j(t+1) \quad = \quad V_{jj}(t) + A_j(t, t+1) + \sum_{k \neq j} V_{kj}(t), \tag{18}$$

where $A_j(t, t+1)$ is the random variable for the arrivals in the interval $[t, t+1)$, with mean $\lambda_j(t)$ and variance $c_{a_j}^2 \lambda(t)$, $V_{jj}(t)$ captures the patients who are staying (not discharging) in unit j, and $V_{k,j}(t)$ captures the patients transferred into unit j from another unit k, for $k, j = 1, \ldots, J$ and $k \neq j$, respectively. Note that for each given unit j,

$$\left(V_{jj}(t), \{V_{j,k}(t)\}_{k \neq j}, X_j(t) - V_{jj}(t) - \sum_{k \neq j} V_{j,k}(t) \right)$$

are multinomial random variables (r.v.) with parameters $X_j(t)$ and corresponding probabilities $1 - \mu_j(t)$, $\{p_{j,k}\mu_j(t)\}_{k \neq j}$, and $(1 - \sum_{k \neq j} p_{j,k})\mu_j(t)$, respectively. That is, the multinomial r.v. corresponds to where the patients currently in service would be the next day – staying in the current unit, transferred to another unit, or discharged (leave the system). The means for the first two set of r.v.'s are

$$E[V_{jj}(t)|X_j(t)] = \big(1 - \mu_j(t)\big)X_j(t), \;\; E[V_{j,k}(t)|X_j(t)] = p_{j,k}\mu_j(t)X_j(t) \;\; \forall k \neq j.$$

Correspondingly, the variances are given by

$$Var[V_{jj}(t)|X_j(t)] = \mu_j(t)\big(1 - \mu_j(t)\big)X_j(t),$$

$$Var[V_{j,k}(t)|X_j(t)] = p_{j,k}\mu_j(t)\big(1 - p_{j,k}\mu_j(t)\big)X_j(t) \;\; \forall k \neq j,$$

and the covariances are given by

$$Cov[V_{jj}(t), V_{j,k}(t)|X_j(t)] = -X_j(t)\big(1-\mu_j(t)\big)p_{j,k}\mu_j(t)\ \forall k \neq j,$$

$$Cov[V_{j,\ell}(t), V_{j,k}(t)|X_j(t)] = -X_j(t)p_{j,\ell}\mu_j(t)p_{j,k}\mu_j(t)\ \forall \ell, k \neq j \text{ and } \ell \neq k.$$

Characterizing the mean. Taking the expectation for both sides of (18), we get the following fluid equation for the mean workload calculation:

$$x_j(t+1) \quad = \quad x_j(t)\big(1-\mu_j(t)\big) + \lambda_j(t) + \sum_{k\neq j} p_{k,j}(t)\cdot x_k(t)\mu_k(t). \tag{19}$$

Compared to the two-station version as in (5), the inflow to unit j now comes from multiple sources: the new patient arrivals $\lambda_j(t)$ and the transfers from all other units $k \neq j$, which equals the sum, for all k's, of $p_{k,j}(t)$ multiplied by the outflow from unit k, $x_k(t)\mu_k(t)$.

Characterizing the variance. Without loss of generality, we take $j = 1$ as an example and the calculation for other stations can be done similarly. Given $X(t)$, using the law of total variance, we get

$$\begin{aligned}
\tilde{v}_{11} &= Var(V_{11}(t)) \\
&= \mathbb{E}[Var(V_{11}(t)|X(t))] + Var[\mathbb{E}(V_{11}(t)|X(t))] \\
&= x_1(t)\mu_1(t)(1-\mu_1(t)) + (1-\mu_1(t))^2 v_1(t),
\end{aligned}$$

where $v_1(t) = Var[X_1(t)]$ is the variance of the workload at t, $X_1(t)$. Similarly, the variance of $V_{k1}(t)$ $(k \neq 1)$, denoted as $\tilde{v}_{k1}$, equals

$$\begin{aligned}
\tilde{v}_{k1} &= Var(V_{k1}(t)) \\
&= \mathbb{E}[Var(V_{k1}(t)|X(t))] + Var[\mathbb{E}(V_{k1}(t)|X(t))] \\
&= x_k(t)p_{k1}\mu_k(t)(1-p_{k1}\mu_k(t)) + (p_{k1}\mu_k(t))^2 v_k(t),
\end{aligned}$$

where $v_k(t) = Var[X_k(t)]$.

Let $cv_{\ell k}(t) = Cov(X_\ell(t), X_k(t))$. For covariance between station 1 and other station k, we get

$$\begin{aligned}
\tilde{cv}_{1k} &= Cov(V_{11}(t), V_{k1}(t)) \\
&= \mathbb{E}[Cov(V_{11}(t), V_{k1}(t)|X(t))] &(20)\\
&\quad + Cov[\mathbb{E}(V_{11}(t)|X(t)), \mathbb{E}(V_{k1}(t)|X(t))]. &(21)
\end{aligned}$$

The second term (21) equals

$$Cov[X_1(t)(1-\mu_1(t)), X_k(t)p_{k1}\mu_k(t)] \quad = \quad (1-\mu_1(t))p_{k1}\mu_k(t)\cdot cv_{1k}(t).$$

The first term (20) equals 0.

Additionally, for covariance between stations $\ell, k \neq 1$, and $\ell \neq k$, we have

$$\begin{aligned} \tilde{cv}_{\ell k} &= Cov(V_{\ell 1}(t), V_{k1}(t)) \\ &= \mathbb{E}[Cov(V_{\ell 1}(t), V_{k1}(t)|X(t))] \qquad (22) \\ &\quad + Cov[\mathbb{E}(V_{\ell 1}(t)|X(t)), \mathbb{E}(V_{k1}(t)|X(t))]. \qquad (23) \end{aligned}$$

Similarly, the first term (22) equals 0, while the second term (23) equals

$$Cov[X_\ell(t)p_{\ell 1}\mu_\ell(t), X_k(t)p_{k1}\mu_k(t)] \quad = \quad p_{\ell 1}\mu_\ell(t)p_{k1}\mu_k(t)\cdot cv_{\ell k}(t).$$

Then, assuming the arrival $A_1(t,t+1)$ is independent of V_{11} and all the V_{k1}'s, we get

$$v_1(t+1) \quad = \quad Var[X_1(t+1)] = \tilde{v}_{11} + c_a^2\lambda_1(t) + \sum_{k\neq 1}\tilde{v}_{k1} + \sum_{\ell\neq k}\tilde{cv}_{\ell k}. \qquad (24)$$

The variance for other stations j, $v_j(t+1) = Var[X_j(t+1)]$ can be calculated in a similar way.

Characterizing the covariance. For the covariance, using the law of total covariance, we have

$$\begin{aligned} cv_{jk}(t+1) &= Cov\big(X_j(t+1), X_k(t+1)\big) \\ &= \mathbb{E}[Cov(X_j(t+1), X_k(t+1)|X(t))] \qquad (25) \\ &\quad + Cov[\mathbb{E}(X_j(t+1)|X(t)), \mathbb{E}(X_k(t+1)|X(t))]. \qquad (26) \end{aligned}$$

For the second term (26), we have

$$\begin{aligned} & Cov[\mathbb{E}(X_j(t+1)|X(t)), \mathbb{E}(X_k(t+1)|X(t))] \\ = \; & Cov\big[(X_j(t)(1-\mu_j(t)) + \sum_{i\neq j} p_{ij}\mu_i(t)X_i(t)), (X_k(t)(1-\mu_k(t)) + \sum_{\ell\neq k} p_{\ell k}\mu_\ell(t)X_\ell(t))\big] \\ = \; & (1-\mu_j(t))(1-\mu_k(t))cv_{jk}(t) + \sum_{\ell\neq k}\big(1-\mu_j(t)\big)p_{\ell k}\mu_\ell(t)cv_{j\ell}(t) \\ & + \sum_{i\neq j}\big(1-\mu_k(t)\big)p_{ij}\mu_i(t)cv_{ki}(t) + \sum_{i\neq j}\sum_{\ell\neq k} p_{ij}\mu_i(t)p_{\ell k}\mu_\ell(t)cv_{i\ell}(t), \end{aligned}$$

with $cv_{ij}(t) = Cov(X_i(t), X_j(t))$ and $cv_{jj} = Var(X_j(t))$. Here, for the first equation, we have used

the fact that the mean arrival rate $\lambda_j(t)$ is a constant and independent of $(X_1(t), X_2(t), \ldots, X_J(t))$.

For the first term (25), we have

$$\begin{aligned}
& \mathbb{E}[Cov(X_j(t+1), X_k(t+1)|X(t))] \\
= \; & \mathbb{E}[Cov\big(V_{jj}(t) + A_j(t,t+1) + \sum_{i\neq j} V_{ij}(t), V_{kk}(t) + A_k(t,t+1) + \sum_{\ell\neq k} V_{\ell k}(t)|X(t)\big)] \\
= \; & \mathbb{E}[Cov\big(V_{jj}(t) + \sum_{i\neq j} V_{ij}(t), V_{kk}(t) + \sum_{\ell\neq k} V_{\ell k}(t)|X(t)\big)] \\
= \; & -x_j(t)(1-\mu_j(t))p_{jk}\mu_j(t) - x_k(t)(1-\mu_k(t))p_{kj}\mu_k(t).
\end{aligned}$$

Here, for the first equation, we have used the fact that arrivals are independent of all the V terms. For the second equation, we have used the fact that conditioning on $X(t)$, $Cov(V_{ij}(t), V_{\ell k}(t)|X(t)) = 0$ except when $i = \ell$, because they are from different independent trials; see similar argument for (7).

B.3 Impact of Network Dynamics on Workload Variance

In this section, we provide a further illustration on the impact of network dynamics on workload variance by writing the single station workload variance as a recursive function for direct comparison with the network variance calculation. For the sake of exposition, let $\mu_j(t) = \mu_j$ for all t. The single-station variance of $X_j(t+1)$ is derived from (4):

$$\sum_{s=0}^{t} \lambda_j(t-s)V_j(s) \;\; = \;\; \lambda_j(t)V(0) + \sum_{s=0}^{t-1} \lambda_j(t-1-s)V_j(s+1),$$

where

$$\begin{aligned}
V_j(s+1) &= \bar{G}_j(s+1) + (c_{a_j}^2 - 1)\bar{G}_j(s+1)^2 \\
&= \bar{G}_j(s)(1-\mu_j) + (c_{a_j}^2 - 1)\bar{G}_j^2(s)(1-\mu_j)^2 \\
&= V_j(s)(1-\mu_j)^2 + \bar{G}_j(s)(1-\mu_j) - \bar{G}_j(s)(1-\mu_j)^2 \\
&= V_j(s)(1-\mu_j)^2 + \bar{G}_j(s)\mu_j(1-\mu_j).
\end{aligned}$$

Thus, the variance of the workload at time $t+1$ from (4) can be written recursively as

$$\begin{aligned}
v_j(t+1) &= \sum_{s=0}^{t-1} \lambda_j(t-1-s)\Big(V(s)(1-\mu_j)^2 + \bar{G}_j(s)\mu_j(1-\mu_j)\Big) + \lambda_j(t)V(0) \\
&= (1-\mu_j)^2 v(t) + \sum_{s=0}^{t-1} \lambda_j(t-1-s)\bar{G}_j(s)\mu_j(1-\mu_j) + \lambda_j(t)c_{a_j}^2 \\
&= (1-\mu_j)^2 v(t) + x_j(t)\mu_j(1-\mu_j) + \lambda_j(t)c_{a_j}^2 \\
&= \tilde{v}_{jj} + \lambda_j(t)c_{a_j}^2, \qquad (27)
\end{aligned}$$

where the first equation follows from the variance calculation for the single-station queue derived in (4). For the second equation, we have used the fact that the workload mean in the single-station queue is calculated by $x_j(t) = \sum_{s=0}^{t-1} \lambda_j(t-1-s)\bar{G}_j(s)$.

Comparing (27) with (9), the impact of the network on the workload variance becomes clear. (27) directly corresponds with the first two terms in (9): the variance from the workload that stays in the unit and the variance of the arrival process. The additional terms in (9) can be interpreted as the isolated contribution of the network to the workload variance in terms of (i) variance from transfers, $\tilde{v}_{21}$, and (ii) correlation between the queue lengths in the two units, $\tilde{cv}_{12}$.