Queueing Models for Patient-flow Dynamics in Inpatient Wards

Hospitals in the US and around the world are struggling with congestion in their inpatient wards (IW), which is largely due to the lack of sufficient resources. Hospital managements are increasingly seeking operational policies to improve patient-flow, and thus reduce related congestions and delays. To this end, queueing models are incorporated into toolkits that are dedicated to help hospitals improve related operations. It is significant that such toolkits build on existing queueing models (such as the $M/M/c$ queue), despite the fact that patient-flow dynamics within the hospital possess unique features that are not captured by those existing models.

In this paper, we identify several important features of hospital queues. We then incorporate these features into appropriate models. Emphasis are given to models that are easy to calibrate and are sufficiently general, so that their analyses and insights are interpretable and applicable in broad settings. Analytically, we characterize long-run regularity of the patient-flow processes, such as the (periodic) time-dependent equilibrium behavior under different operational policies. We also develop accurate approximations to facilitate efficient performance evaluation and optimization.

1. Main Characteristics of IW Queues

Since patient-flow dynamics are highly complex, an effective model must be relatively parsimonious, “distilling” only the most salient characteristics of the related queueing processes. We now elaborate on the three main features that are incorporated in our model.
**Periodic Discharge Decisions.** Unlike a typical queueing system, in which customers depart the server immediately when their service terminates, a recovered patient does not leave the bed until a physician examines and approves her release. In large hospitals, most patients are examined during daily inspection rounds that take place in the morning hours. Thus, patients may occupy their beds for relatively long periods of time *after their recovery*, while waiting to be examined by a physician.

**Discharge Delays.** The departures of a patient who were judged to be recovered is further delayed due to several reasons (e.g., paperwork, need for transportation arrangements, coaching by professionals, etc.), and occur in batches, several hours after the morning inspection round has ended. These delays lead to a relatively low variability of the discharge times, as most departures tend to be concentrated at a narrow time interval that is highly predictable.

**Time Varying Arrivals.** Like most service systems in practice, the arrival rate (of bed requests) to the IW is time-varying. However, unlike many other service systems for which stationary analysis can be effective despite having time-dependent arrival rates, the time-dependent queuing dynamics corresponding to patient-flow must be incorporated into a queueing model for the IW. This is because the arrival rate of bed-request changes considerably over a day, while the average hospitalization period is several-days long.

### 2. Main Contribution

In this paper, we propose a new queueing model for IWs, taking into account the unique features associated with patient-flow dynamics. The model is easy to calibrate and robust. Currently, the model has been incorporated into a toolkit that
is employed by a large teaching hospital in the US in a capacity-planning project. By analyzing the model, we make the following contributions.

(I) We show that the effective service-time of a patient depends on her arrival time, so that a “service rate” does not exist. Nevertheless, we characterize the maximum effective service rate as an explicit function of the model primitives, and in particular, of bed blocking that is due to the special discharge process. In turn, this allows us to quantify the maximal throughput of the IW, namely, the maximum number of arrivals per day that the IW can handle without making changes to its treatment or discharge policies. Even though our system is not stationary, we prove that it converges to a periodic steady state provided the average number of daily arrivals is smaller than the maximum throughput.

(II) Since the stochastic dynamics of our model are analytically intractable, we develop an tractable fluid model to approximate those complex patient-flow dynamics. We then prove that, just like the underlying stochastic system, the fluid approximation possesses a unique periodic steady state to which it converges for all initial conditions. This convergence simplifies long-run analysis considerably, because long-run computations reduce to the analysis of a single period (the stationary one). We demonstrate how this “stationary-type” analysis facilitates the computations of key performance measures. In particular, one can easily compute the impact of changing the time of the inspection round or the discharge delay on the queue length, waiting times, or the proportion of patients that wait more than a given time period for an IW bed.