We model the patient-flow dynamics associated with Inpatient Wards (IW) in large hospitals. Our model aims to capture the most salient aspects of the “service process” (hospitalization period) in the IW, that are unique to this setting, and their affect on key performance measures. In particular, since discharging a patient requires a physician’s approval, patients typically occupy their beds for long time periods after their “service” (treatment) is complete, even though they are medically ready to leave the bed. In addition, most departures from the IW tend to be highly concentrated in a short time period each day, that is several hours after the discharge decisions have been made. (We refer to this latter phenomenon as discharge delays.) Therefore, patients occupy their beds for long time periods after their service has ended, and service times are not independent and identically distributed across the patients. These features, in addition to the non-stationarity of the arrival process of bed requests, render stochastic analysis prohibitively hard. Our model is intended to facilitate strategic decision making, such as the number of IW beds that are needed, as well as the long-run costs and benefits of reducing delays or changing the discharge decision process. In this paper we quantify the effect of discharge delays on the effective traffic intensity, and characterize the stability condition of the model. To this end, we quantify the traffic intensity to the IW and its stability condition, and employ simple fluid models to approximate key performance measures.