Estimating Queue Externalities in the Emergency Department

Emergency Departments (EDs) operate according to the principle that high-acuity patients should be treated as quickly as possible, ahead of low-acuity patients. Prioritizing treatment for high-acuity patients saves both lives and money, as even small delays in a high-acuity patient’s treatment can negatively affect mortality, recovery times and associated medical costs. In response, many Operations Management studies choose to model EDs and other hospital facilities as preemptive priority queues with multiple acuity classes (Fomundam and Herrmann 2007; Green 2006). Such models frequently assume that a high-acuity patient’s treatment will only be delayed if a facility is currently overloaded with other high-acuity patients. In practice, however, ED crowding by low-acuity patients correlates with significantly increased wait times for even high-acuity patients, indicating that preemptive priority models oversimplify the reality on the ground at many EDs (McCarthy et al. 2009).

In this paper, we combine insights from queueing theory with econometrics to estimate the amount by which low-acuity patients can increase the wait time to start of treatment for high-acuity patients. Specifically, we estimate a lower bound on the “high-acuity queue externality” (the marginal change in high-acuity patients’ wait times caused by an arriving low-acuity patient) at four hospitals. We also leverage existing estimates of how waiting impacts mortality, hospital stays, and hospital costs in order to find that low-acuity arrivals may substantially increase mortality, hospital stays, and hospital costs, in one ED under study. At that ED, we also validate our findings using an on-site experiment. Based on both data analysis and on-site interviews with hospital staff, we suggest managerial interventions EDs might use to mitigate these effects.

Concerning our estimation procedure, we model the queue externality as a sum of nonnegative random variables. Each element of the sum corresponds to the amount by which a given low-acuity patient increases one high-acuity patient’s wait time. The distribution of each random variable may depend on when the two patients visited the ED relative to one another: for example, a low-acuity patient has a larger potential impact on the wait time of a high-acuity patient who arrived 5 minutes later, compared to a high-acuity patient who arrived several days later. The sum of these random variables for a given
low-acuity patient across a subset of high-acuity patients represents the total impact the low-acuity patient has on those high-acuity patients’ wait times.

To estimate the high-acuity queue externality empirically, we fit a regression on low-acuity patient data where the dependent variable is the log-sum of the wait times of all high-acuity patients whose wait times may have been impacted by that low-acuity patient. Most importantly, we include the number of low-acuity patients in the ED as a covariate, because exponentiating its coefficient gives us the average percent change in the sum of high-acuity wait times associated with the arrival of one more low-acuity patient to the ED. This number in turn allows us to quantify the externality.

We implement the above analysis using data from four hospitals, which we refer to as SMMC, Hospital 1, Hospital 2 and Hospital 3 (the latter 3 having requested anonymity). At SMMC, we take advantage of an on-site experiment to develop instrumental variables that validate our results. We find that the average lower bound on the high-acuity externality is approximately 11.3 minutes at SMMC, 2.0 minutes at Hospital 1, 20.5 minutes at Hospital 2 and 9.5 minutes at Hospital 3. Combining these estimates with average arrival rates and the queue length in the ED, we estimate that, on average, a high-acuity patient waits for at least 14 additional minutes on account of low-acuity patients at SMMC, 7 additional minutes at Hospital 1, 9 additional minutes at Hospital 2, and 5 additional minutes at Hospital 3. These results in turn suggest that 10% to 100% of high-acuity patients’ wait times to treatment may arise due to low-acuity congestion.

Concerning the implications of this work, our empirical evidence for the high-acuity externality indicates that high-acuity treatment is not always perfectly preemptive. Moreover, because even small increases in wait times to treatment can translate to increased mortality, longer lengths of stay, greater repercussions for longterm health, and higher health care costs, the high-acuity externality may have serious repercussions for patient outcomes [Cardosos et al., 2011; Singer et al., 2011]. For example, we estimate that the decrease in low-acuity ED utilization caused by the Affordable Care Act has saved over 14 lives annually in SMMC’s county alone, solely due to the corresponding decrease in the high-acuity externality.
In order to mitigate the high-acuity externality, we propose two managerial interventions. First, to improve high-acuity processing rates, we recommend an intervention trialled at SMMC which involves beginning tests for certain low-acuity patients while they are still in their chairs, without assigning them to beds. At SMMC we estimate that enacting this policy is on average equivalent to adding one additional bed to the ED, and consequently decreases high-acuity wait times by as much as 26%. Second, we introduce ongoing work to design a policy that will encourage low-acuity patients with minor conditions to leave without being seen at times when the high-acuity externality is large. Such a policy uses mechanisms such as the triage room wait time screen to increase the likelihood that non-urgent patients will renege and seek primary care.

References


