BMJ Open

Changes to physician processing times in response to clinic congestion and patient punctuality: a retrospective study

Chester G Chambers,[1] Maqbool Dada,[1] Shereef Elnahal,[2] Stephanie Terezakis,[2] Theodore DeWeese,[2] Joseph Herman,[2] Kayode A Williams[3]

To cite: Chambers CG, Dada M, Elnahal S, *et al*. Changes to physician processing times in response to clinic congestion and patient punctuality: a retrospective study. *BMJ Open* 2016;**6**:e011730. doi:10.1136/bmjopen-2016-011730

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-011730).

All work is attributed to this department and institution and support was provided solely from institutional and/or departmental sources.

Received 9 March 2016
Revised 30 August 2016
Accepted 16 September 2016

For numbered affiliations see end of article.

Correspondence to
Dr Kayode A Williams;
kwilli64@jhmi.edu

ABSTRACT

Objectives: We examine interactions among 3 factors that affect patient waits and use of overtime in outpatient clinics: clinic congestion, patient punctuality and physician processing rates. We hypothesise that the first 2 factors affect physician processing rates, and this adaptive physician behaviour serves to reduce waiting times and the use of overtime.

Setting: 2 urban academic clinics and an affiliated suburban clinic in metropolitan Baltimore, Maryland, USA.

Participants: Appointment times, patient arrival times, start of service and physician processing times were collected for 105 visits at a low-volume suburban clinic 1, 264 visits at a medium-volume academic clinic 2 and 22 266 visits at a high-volume academic clinic 3 over 3 distinct spans of time.

Intervention: Data from the first clinic were previously used to document an intervention to influence patient punctuality. This included a policy that tardy patients were rescheduled.

Primary and secondary outcome measures: Clinicians' processing times were gathered, conditioned on whether the patient or clinician was tardy to test the first hypothesis. Probability distributions of patient unpunctuality were developed preintervention and postintervention for the clinic in which the intervention took place and these data were used to seed a discrete-event simulation.

Results: Average physician processing times differ conditioned on tardiness at clinic 1 with $p=0.03$, at clinic 2 with $p=10^{-5}$ and at clinic 3 with $p=10^{-7}$. Within the simulation, the adaptive physician behaviour degrades system performance by increasing waiting times, probability of overtime and the average amount of overtime used. Each of these changes is significant at the $p<0.01$ level.

Conclusions: Processing times differed for patients in different states in all 3 settings studied. When present, this can be verified using data commonly collected. Ignoring these behaviours leads to faulty conclusions about the efficacy of efforts to improve clinic flow.

Strengths and limitations of this study

- This is the first study to examine whether clinicians reduce their face times with patients when running behind schedule.
- This work uses a simulation model validated in our earlier published work to relate this adaptive behaviour to clinic performance.
- Using simulation we can identify the effects of changes in patient behaviour and physician behaviour in the same system.
- Our analysis shows that not recognising clinicians' responses to falling behind can lead to a biased assessment of the impact of changes in other variables or interventions.
- Results are limited since our analysis is based on clinics within a single metropolitan area and a single medical system.

INTRODUCTION

Over 50 years have elapsed since the seminal work of White and Pike[1] that used discrete-event simulation (DES) to explore linkages between punctuality for patients and physicians in outpatient clinics and metrics of clinic performance, including waiting times and use of overtime. As pressure to serve higher volumes of patients at lower costs has continued to mount, clinic managers have responded by continuing to invest in the use of DES and other operational research techniques[2–6] to evaluate alternative approaches to address these issues. Although the use of DES as a tool to search for a better scheduling algorithm that will improve patient flow has a long history in the research literature,[7–10] one less-studied alternative is to find methods that change the behaviour of key stakeholders in ways that reduce variability or streamline patient flows. As an exemplar, we recently reported on an

intervention in which a rescheduling policy was used for tardy patients.[11] Implementation of this policy induced a change in patient behaviour that included fewer late arrivals and an increase in average earliness. The effect reduced variability in the patient arrival process. That work used a DES model to help quantify the potential impacts of such changes. Clearly physician behaviour will also have an effect on system performance and can be modelled in a parallel fashion.

Our current work is an effort to extend the existing literature by quantifying one type of adaptive behaviour involving physicians in outpatient clinics. Specifically, we performed a retrospective analysis of flow-related data at several specialty appointment-based clinics to learn whether physicians adjust processing rates based on congestion in the clinic. While such behaviour has been documented in prior research on other service delivery systems,[12 13] it has not been considered in efforts to model the performance of outpatient clinics. Thus, the first goal of this study was to determine whether processing times depended on the status of the patient in the system, where 'status' as formally explained later is defined based on whether the patient arrived on time and whether the patient was seen by the appointment time. The second goal was to demonstrate that for cases in which this holds, it had a material impact on clinic performance and should be included when evaluating interventions to improve patient flow.

MATERIALS AND METHODS

We reviewed data collected within three ambulatory services. Clinic 1 was a pain management clinic with one attending physician. Institutional Review Board (IRB) exemption was obtained from Johns Hopkins School of Medicine since these data sets were collected as part of quality improvement or audit-based studies. Data were collected from February 2008 through July 2009 on paper forms that were attached to each patient record on check-in for an appointment. This clinic was run as a private practice that was part of the Johns Hopkins medical system in the Baltimore metropolitan area. A more detailed description of this setting, the data collected and the DES model used is provided in Williams *et al*.[11 14] Clinic 2 was a medium-volume pain management clinic that was also part of the Johns Hopkins medical system. Data were collected from February to March 2010 by paid observers. The primary differences are that this clinic operated at a higher volume and included a teaching mission. A typical clinic session involved one attending physician working in concert with three residents/fellows serving patients on a fixed schedule. The process flow in this setting was made more complex by the teaching mission of the hospital. A detailed description of the setting, data collected and simulation developed is provided in Williams *et al*.[14 15]

Clinic 3 was a high-volume radiation oncology service, which was part of the same medical system. Data from October 2014 through March 2015 were retrieved directly from relevant information technology systems. This clinic accommodated multiple attending physicians who worked simultaneously, blended with a collection of residents, fellows, physicians and the nursing staff. Detailed description of this setting was provided in Elnahal *et al*.[16] We note that for all three of these clinics, the data collected were part of efforts focused on improving performance metrics, including patient waiting times, throughput, and use of overtime, and not to test whether processing rates differ depending on congestion. Consequently, a natural experiment emerged in that it was only after the data were in hand that we uncovered the phenomenon in question.

We categorised patients into three collectively exhaustive and mutually exclusive groups: group A patients were those who arrived and were placed in the examination room before their scheduled appointment time; group B patients were those who arrived before their appointment times, but were placed in the examination room after their appointment time, indicating that the clinic was congested; and group C patients were those who were tardy, meaning that they arrived after their appointment time.

To link patient status to system congestion, we proposed a simple categorisation scheme based on the patient arrival time (w), appointment time (x), the time that the patient entered an examination room (y) and the time that the patient exited the system (z). Generally, when a patient arrived for a scheduled visit, the arrival time was recorded when the patient signed in to the clinic. We compared each of these sign-in times (w) directly to the appointment time (x) to define the patient unpunctuality as x−w. If w was greater than x, the patient was deemed tardy (group C); otherwise the patient was deemed early. In all three settings, data collection included a time stamp indicating when the patient was escorted to an examination room (y). This time was easily compared with the appointment time. If both w and y were less than x, the patient was in group A. If the sign-in time was earlier than the appointment time (w<x), but the time in the room was after the appointment time (x<y), the patient was in group B.

Process time was defined as the span from when the patient entered the examination room (y) until the time that the patient exited the system (z). It is commonly assumed that the distribution of process time will be the same regardless of which group the patient is in. When modelling such systems, this assumption is natural because it is unknown before the patient arrives. We note that this is different than the common assumption that processing time distributions differ based on other characteristics. For example, in many settings, new patients have longer (on average) process times than do return patients. However, this visit type is known before the patient arrives and is a separate phenomenon.

Analysis: simulation models

We previously developed a detailed DES model of clinic 1 as well as detailed information on patient punctuality before and after an intervention to affect patient behaviour.[11] Details of the 11-patient schedule and input values for the DES were also provided there. Considering patient punctuality as one dimension, we defined the preintervention distribution of patient punctuality as well as the postintervention distribution of patient punctuality. Parameters of these distributions were provided along with the description of the DES. Considering treatment times as an orthogonal dimension, we have the distributions of treatment times used in that work. These specifications of process time distributions assumed that the system did not adapt to patient status; thus, the processing times were viewed as one homogeneous set. We refer to these as the 'pooled' processing times. Retrospective analysis of the same data allowed us to group observations based on patient status and the clinician's adaptation to that status. We refer to these collections of processing times as 'adaptive'. This method yielded four distinct settings with two distinct distributions of patient punctuality: preintervention and postintervention, along with two distinct sets of distributions of processing times: pooled and adaptive.

Analysis: metrics of interest

The cost elements related to the performance of outpatient clinics that are relevant to this analysis include those associated with patient waits and clinic overtime. Waiting time is one of the most common complaints about outpatient clinics.[17] Given variable processing times, waiting times can be eliminated only if excess slack is built into the schedule. This comes at the expense of increasing the average session duration, which we refer to as MAKESPAN. Increasing MAKESPAN leads to exceeding the targeted end of the clinic session. We will refer to the amount by which the end of the clinic session exceeds this target as overtime.

We used the DES model to compute values of average waiting times (WAIT) as well as the average duration between the patient's appointment time and the start of service (DELAY). We report both results because it is often argued that appointments create expectations in the mind of the customer such that waits after the appointment time are viewed differently from waits before the appointment time.[18] Note that each patient's wait includes the delay if any. For example, waiting time can be positive, but if it all takes place before the appointment time, the patient delay is still 0. We refer to the duration between a starting point and exit from the system as flow time. When we use arrival time as the starting point, we refer to the average flow time as FT-ARR. When we use the appointment time as the starting point, we label this as FT-APPT. In addition, we report the proportion of patients who experience a positive value of delay (proportion delayed). To explore the effect of a system change on the three patient groups, we also report these patient-related metrics for each of the three groups in each setting. To explain the overtime-related costs, we report the average session duration (MAKESPAN), the number of patients on the 11-patient schedule exiting the system by 12:00 (Comp by 12:00), the probability of having a positive level of overtime (Prob OT), the average amount of session overtime (Ave OT) and the average amount of session overtime when it is positive (Ave POT).

RESULTS

Processing times by patient group

Table 1 shows characteristics of process times for the three patient groups in each of the three clinic settings. For clinic 1, the average processing times and SEs for groups A, B and C were 38.31 (3.21), 26.23 (2.23) and 29.50 (3.47) min, respectively. For clinic 2, these values were 65.59 (2.24), 53.53 (1.97) and 50.91 (3.11) min, respectively. For clinic 3, the average processing times for the three groups were 47.15 (0.81), 17.59 (0.16) and 47.90 (1.59) min, respectively. When we tested the hypothesis that average processing times were the same for groups A, B and C using an F test as described in Brunk,[19] the resulting p values for the three tests were 0.03 for clinic 1, 10^{-5} for clinic 2 and 10^{-7} for clinic 3. Thus, in each case, we rejected the hypothesis that all three means are equal in favour of the alternative that the means are not equal.

Given that a patient shows up on time, we next tested the hypothesis that average processing times was greater for group B when compared with the corresponding group A. (The difference in sample means for group C relative to its corresponding groups A and B showed no

Table 1 Processing times pooled and by group for each clinic

	Clinic 1	Clinic 2	Clinic 3
PROCESS (min)	34.65	57.98	35.22
n/SE—PROCESS	105/2.30	264/1.42	22 266/0.44
A time (min)	38.31	65.59	47.15
n/SE A time	71/3.21	116/2.24	10 352/0.81
Prop A (%)	67.62	43.94	46.49
A time (%)	110.56	113.13	133.87
B time (min)	26.23	53.53	17.59
n/SE B time	26/2.23	101/1.97	9058/0.16
Prop B (%)	24.76	38.26	40.68
B time (%)	75.70	92.32	54.04
C time (min)	29.50	50.91	47.90
n/SE C time	8/3.47	47/3.11	2855/1.59
Prop C (%)	7.62	17.8	12.82
C time (%)	85.14	87.81	136.00

A time, average processing time for patients in group A; A time (%), ratio of average processing time for group A compared with the global average written as a percentage; n/SE—PROCESS, number of observations and SE of processing time; n/SE A time, SE of A time; PROCESS, average value of exit time minus time the patient enters the examination room; Prop A, percentage of population in group A.

clear ranking. And even when the difference was statistically significant; eg, in clinic 3 when comparing groups A and C, there was no practical significance since the means are relatively close to each other.) This was done using several approaches. Using a one-sided t-test assuming that variances differ between groups in each setting, the resulting p values were $3.15e^{-08}$ for clinic 1, $4.84e^{-06}$ for clinic 2 and $1.1e^{-16}$ for clinic 3. Using a one-sided Tukey Honest Significant Difference (HSD) test the resulting p values were 0.03, $3e^{-5}$ and $<10^{-7}$. A Kolmogorov-Smirnov test produced corresponding p values of 0.03, $6.1e^{-4}$ and $2.2e^{-16}$. Finally the Mann-Whitney-Wilcoxon rank-sum test, which does not assume normality of the data sets, produced p values of $8.9e^{-3}$, $1.8e^{-5}$ and $2.2e^{-16}$, respectively. Thus, we conclude that the process times differed by patient status. Given this conclusion, the remaining phase of this work was to explore the implications of this fact. We approached this objective by revisiting the DES model used in the analysis of clinic 1 that was described in our previous work.[11] The details of this model are described in our previous work. For the reader's convenience, we replicate several tables and figures used to describe the simulation in the online supplementary appendix.

Tables 2 and 3 report results from simulation models of the four scenarios using the relative processing times recorded in clinic 1. For ease of exposition, we refer to the case of adaptive processing times prior to the intervention as the base case. In this adaptive system, we can also say that processing times for patients in group A are 1.22 times the average processing time, processing times for group B are 0.84 times the average and processing times for group C are 0.94 times the average. All metrics reported from the DES were averages over 10 000 simulated sessions, and all differences were significant at p<0.001 level.

When comparing preintervention and postintervention levels of patient punctuality in the adaptive scenarios, we found that WAIT increased from 17.28 to 18.26 min and FT-ARR increased from 76.36 to 78.60 min, but FT-APPT decreased from 59.62 to 57.85 min and DELAY decreased from 10.84 to 9.15 min. Thus, average waiting times increased but more of the wait was experienced before the appointment time. For the clinic as a system, MAKESPAN decreased from 261.92 to 260.47 min, the number of patients who cleared the system by 12:00 increased from 9.32 to 9.39 patients, proportion delayed decreased from 48.34% to 39.78%, Prob OT decreased from 76.06% to 71.18%, expected OT decreased from 18.82 to 17.10 min and expected overtime when positive decreased from 24.74 to 24.02. Thus, as in our previous work,[11] a savings of ~1.72 min of overtime came at the expense of 10.78 extra minutes in the clinic on average for the 11 patients as a group.

It might seem unintuitive that an intervention that resulted in virtually all patients changing behaviour by arriving earlier relative to appointment times improved overtime costs only modestly. Some reconciliation is possible by looking at the disaggregated results in table 3. When we compared the performance metrics in the adaptive scenarios preintervention and postintervention, we found that the proportion of group A patients increased substantially (51.65% vs 60.22%), as did the proportion of group B patients (33.94% vs 39.13%). Since group A patients tend to have longer process times, the dynamics of the system change subtly. Longer process times increase the likelihood of congestion, which explains the higher proportion of group B patients. Treating group B patients tends to take less time, so it tends to bring the clinician back on schedule. This feedback is remarkably stable because the weighted average process time preintervention was 33 min and postintervention was 33.5 min. Since system usage did not change, having earlier arrivals had little impact on system performance.

In contrast, in our previous work,[11] service times were drawn from a pooled distribution. When we compared

Table 2 Performance metrics under four key scenarios

	Preintervention, adaptive	Postintervention, adaptive	Preintervention, pooled	Postintervention, pooled
WAIT (min)	17.28	18.26	15.90	16.90
DELAY (min)	10.84	9.15	9.59	7.86
FT-ARR (min)	76.36	78.60	70.19	71.21
FT-APPT (min)	59.62	57.85	53.57	50.49
Proportion delayed (%)	48.34	39.78	48.29	39.70
MAKESPAN (min)	261.92	260.47	250.39	244.50
Comp by 12:00 (%)	9.32	9.39	9.91	10.12
Prob OT (%)	76.06	71.18	61.22	49.40
Exp OT (min)	18.82	17.10	9.72	6.32
Ave POT (min)	24.74	24.02	15.88	12.79

Ave POT, average overtime when it is strictly positive; Comp by 12:00, percentage of 10 000 sessions in which last patient leaves system before 12:00; DELAY, average gap between appointment time and entrance to examination room; Exp OT, average overtime used across all sessions; FT-APPT, gap between appointment time and exit from system; FT-ARR, gap between patient arrival and exit from system; MAKESPAN, the time span from the start of the clinic session to the completion of the last patient on the schedule; Prob OT, proportion of sessions in which MAKESPAN exceeded 240 min; Proportion delayed, proportion of patients whose start of service exceeds appointment time; WAIT, average gap between patient arrival and entrance to examination room.

Table 3 Performance metrics for each patient group

	Preintervention, adaptive	Postintervention, adaptive	Preintervention, pooled	Postintervention, pooled
Prop group A (%)	51.65	60.22	51.71	60.30
WAIT A (min)	6.44	6.30	6.36	6.22
DELAY A (min)	NA	NA	NA	NA
FT-ARR A (min)	75.54	76.02	66.07	65.10
FT-APPT A (min)	48.91	50.16	39.50	39.25
Prop group B (%)	33.94	39.13	33.82	39.05
WAIT B (min)	41.40	31.96	27.71	28.31
DELAY B (min)	19.63	19.94	16.44	16.63
FT-ARR B (min)	77.78	77.98	75.59	75.55
FT-APPT B (min)	65.91	65.94	64.20	63.86
Prop group C (%)	14.40	0.65	14.47	0.65
WAIT C (min)	11.62	0.96	10.06	0.85
DELAY C (min)	11.62	0.96	10.06	0.85
FT-ARR C (min)	67.58	67.02	62.96	61.97
FT-APPT C (min)	76.68	67.77	72.13	62.71

DELAY A, average delay for all patients found to be in group A; FT-APPT A, average gap between appointment time and exit time for patients found to be in group A; FT-ARR A, average gap between arrival time and exit time for patients found to be in group A; NA, not applicable; Prop group A, proportion of patients found to be in group A; WAIT A, average waiting time for all patients found to be in group A.

preintervention and postintervention levels of patient punctuality in the pooled scenarios, we found that WAIT increased from 15.90 to 16.90 min and FT-ARR increased from 70.19 to 71.21 min, but FT-APPT decreased from 53.57 to 50.49 min. Again, average waiting times increased but more of the waiting time was experienced before the appointment time. Average MAKESPAN decreased from 250.39 to 244.50 min, and the number of patients who cleared the system by 12:00 increased from 9.91 to 10.12. The proportion of patients delayed fell from 48.29% to 39.70%, the probability of OT fell from 61.22% to 49.40%, the expected OT fell from 9.72 to 6.32 min and the expected OT when positive fell from 15.88 to 12.79. Thus, a reduction of expected overtime by about 3.4 min came at the expense of about 11.2 additional patient minutes in the clinic.

Since pooling distributions removed one dimension of variability, we should expect to see overall better performance in the pooled case. For example, if we compare the adaptive scenario to the pooled scenario, including the preintervention levels of patient punctuality, we see that WAIT falls from 17.28 to 15.90 min; DELAY falls from 10.84 to 9.59 min; MAKESPAN falls from 261.92 to 250.39 min; FT-ARR falls from 76.36 to 70.19 min, FT-APPT falls from 59.62 to 53.57 min; patients clearing the system by 12:00 rises from 9.32 to 9.91; proportion delayed decreases from 48.34% to 48.29%; Prob OT falls from 76.06 to 61.22; and Ave OT falls from 18.82 to 9.72 min and Ave POT falls from 24.74 to 15.88 min. Analogous results held when we considered the postintervention levels of patient punctuality. However, the relative changes between preintervention and postintervention settings were in the same direction and had similar magnitudes. Thus, we concluded that if clinicians were to change behaviour by using a pooled processing distribution, which has the same mean but lower variability, it would save ~9.1 min of overtime cost per session, while also reducing patient time in the clinic by ~67.9 min.

Just as using the pooled system instead of the adaptive system as the representation of the actual clinic would result in underestimating overtime costs, it would also overstate the relative disparity between group A and B patients. By examining FT-ARR in the preintervention pooled column of table 3, we see that on average, group A patients exit the clinic 66.07 min after their arrival. Parallel values for groups B and C are 75.59 and 62.96 min, respectively. These results show that tardy patients spend the least amount of time in the system, whereas on average, a group B patient spends 9.5 min more than a group A patient and 12.63 min more than a group C patient. In contrast, in the adaptive system, FT-ARR values are longer but the disparity is less pronounced (75.54, 77.78, 67.58 min). Thus, we concluded that the relative disadvantage of being a group B patient is attenuated in the adaptive system, but the average flow time is longer.

Though we may posit that pooling processing times so that all patients are treated similarly is an important form of changing physician behaviour, it is not the only possible change. In tables 4 and 5, we present metrics from a scenario in which all patients are treated like group B patients. Clearly this is possible, as it is done in the adaptive system based on patient status. Such a change would reduce both the mean and variability in processing times. Standard results from queueing theory suggest that the benefits would be even more pronounced relative to the adaptive base case. Putting aside the impact of reduced time with the doctor, treating all patients as if they are from group B creates a benchmark against which proposed behavioural adaptations can be

Table 4 Performance metrics with all patients treated as group B

	Preintervention, all B duration	Postintervention, all B duration
WAIT (min)	11.42	12.01
DELAY (min)	5.95	4.01
FT-ARR (min)	57.01	57.45
FT-APPT (min)	40.39	36.75
Proportion delayed	43.59	33.92
MAKESPAN (min)	231.89	224.70
Comp by 12:00 (%)	10.60	10.78
Prob OT (%)	30.63	17.38
Exp OT (min)	1.31	0.43
Ave POT (min)	4.28	2.47

Ave POT, average overtime when it is strictly positive; Comp by 12:00, percentage of 10 000 sessions in which last patient leaves system before 12:00; DELAY, average gap between appointment time and entrance to examination room; Exp OT, average overtime used across all sessions; FT-APPT, gap between appointment time and exit from system; FT-ARR, gap between patient arrival and exit from system; MAKESPAN, the time span from the start of the clinic session to the completion of the last patient on the schedule; Prob OT, proportion of sessions in which MAKESPAN exceeded 240 min; Proportion delayed, proportion of patients whose start of service exceeds appointment time; WAIT, average gap between patient arrival and entrance to examination room.

Table 5 Performance metrics for each patient group with all patients treated as group B

	Preintervention, all B duration	Postintervention, all B duration
Prop group A (%)	56.41	66.07
WAIT A (min)	6.61	6.51
DELAY A (min)	NA	NA
FT-ARR A (min)	54.64	53.60
FT-APPT A (min)	28.10	27.69
Prop group B (%)	29.12	33.28
WAIT B (min)	18.51	18.77
DELAY B (min)	9.71	9.72
FT-ARR B (min)	60.66	60.65
FT-APPT B (min)	51.69	51.58
Prop group C (%)	14.47	0.64
WAIT C (min)	7.22	0.53
DELAY C (min)	7.22	0.53
FT-ARR C (min)	52.30	51.31
FT-APPT C (min)	61.39	52.04

DELAY A, average delay for all patients found to be in group A; FT-APPT A, average gap between appointment time and exit time for patients found to be in group A; FT-ARR A, average gap between arrival time and exit time for patients found to be in group A; NA, not applicable; Prop group A, proportion of patients found to be in group A; WAIT A, average waiting time for all patients found to be in group A.

compared. As can be readily seen from the tables, the improvement in performance metrics becomes more pronounced. For example, relative to the base case, expected overtime would fall from 18.82 to 9.72 min in the pooled system and to 1.31 min in this benchmark case.

Interestingly for all three sets of scenarios, the incremental benefit of the patient punctuality intervention resulted in an additional reduction in overtime of <2 min, suggesting that influencing clinician behaviour has a substantially bigger effect than does improving patient punctuality.

DISCUSSION

Several studies, including our own,[11] document some form of adaptive behaviour in patient populations. In particular, that work looked at an intervention meant to improve patient punctuality and thereby change the arrival pattern in a way that would reduce its variability. This reduction in variability has the effect of improving process flow. In other words, as patients become more punctual, the system improves. On the other hand, the adaptation highlighted in our current work, involves clinicians. Specifically, it appears that in response to system congestion, clinicians change processing rates as a function of patient status. This adaptation has negative consequences for the system because a new source of variability is introduced. This factor degrades system performance along every dimension that we measured. Ironically, longer service times for group A increase the proportion of group B patients, who have shorter service times, yet this is not enough to compensate for the higher usage of resources by patients in group A. In other words, ignoring the adaptive system behaviour undercounts the service time of group A patients, underestimating the congestion in the system. The higher congestion degrades system performance. Finally, as a benchmark, we looked at the setting in which all patients are treated as though they are in group B. In this setting, both the mean processing times and the variability of those times were reduced, yielding a dramatic improvement in system performance.

One key message of this work is that ignoring sources of variability can lead to misleading results. When we compare flow metrics using the two distributions of patient punctuality, the improvement is consistent if we assume that the system has either pooled processing times or the adaptive processing times, as long as we are consistent across the two cases. Thus, our previous results[11] hold. However, ignoring the adaptive system behaviour implies that work understated the magnitude of the original problem. We suspect that the same issue is present in many studies that examine efforts to improve patient flow but ignore the adaptive behaviour of physicians. It feels natural to assume that adaptive behaviour makes the system flow more smoothly. Our results suggest that this general conclusion is incomplete. If the adaptation increases variability, it includes unintended consequences that may make things worse.

After considering three clinics in some detail, we found that in all three settings, the distributions of processing times differed for different groups of patients. We should note that these three systems appear to

operationalise this accommodation in different ways. If the service involves a single physician, this accommodation appears to be in the form of the attending 'going faster' when the clinic is behind schedule. This may be seen as a natural response to falling behind. Clinics 2 and 3 add a teaching mission, which alters process flow because it adds steps related to interactions between trainees and patients as well as interactions between trainees and the attending. The fact that the differences between processing times for group A and B patients is greater in clinics 2 and 3 than in clinic 1 suggests that the changes in processing rates may be related to how the trainee is managed in the process. For example, one natural explanation for why the system would move faster when behind schedule in the academic setting could be that the trainee may be removed from the process for specific patients or the presence of the trainee may allow parallel processing of patients. This possibility has been discussed in detail previously.[14 15]

Additional work is needed to verify the details of the adaptation mechanism along with efforts to discern whether it is present in other clinical environments, but this work is the first step in demonstrating that it does indeed occur and that it does have an impact on system congestion. We also believe that its inclusion in models of delivery systems may prove to be a significant step forward in making approaches based on operational research techniques more salient and reliable in healthcare settings.

CONCLUSIONS

In some outpatient settings, processing times are related to system congestion. This can be verified using data commonly collected to report clinic performance. If this behaviour is present, it is useful to take it into account when describing patient flow and interpreting the efficacy of actions designed to improve clinic performance. Both researchers and practitioners could benefit by taking steps to revisit existing models and build new models that provide a more comprehensive depiction of system performance.

Author affiliations
[1]Johns Hopkins Carey Business School, Armstrong Institute for Patient Safety and Quality, Baltimore, Maryland, USA
[2]Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins School of Medicine, Baltimore, Maryland, USA
[3]Department of Anesthesiology and Critical Care Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA

Contributors CGC helped design the study, conduct the study, data analysis, prepare the manuscript and approved the final manuscript. MD helped design the study, conduct the study, data analysis, prepare the manuscript and approved the final manuscript. SE helped with data collection and approved the final manuscript. ST helped with data collection and approved the final manuscript. TD helped conduct the study, review the data, prepare the manuscript and approved the final manuscript. JH helped conduct the study, prepare the manuscript and approved the final manuscript. KAW helped design the study, conduct the study, prepare the manuscript and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data are available via email with the corresponding author.

REFERENCES

1. White MJB, Pike MC. Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Med Care* 1964;2:133–45.
2. Fetter RB, Thompson JD. The simulation of hospital systems. *Operations Res* 1965;13:689–711.
3. Clymer JR. *Simulation-based engineering of complex systems*. Hoboken, NJ: John Wiley & Sons, 2009.
4. Jun JB, Jacobson SH, Swisher JR. Application of discrete-event simulation in health care clinics: a survey. *J Operational Res Soc* 1999;50:109–23.
5. Mahachek AR. An introduction to patient flow simulation for health-care managers. *J Soc Health Syst* 1992;3:73–81.
6. Benneyan JC. An introduction to using computer simulation in healthcare: patient wait case study. *J Soc Health Syst* 1997;5:1–15.
7. Rohleder TR, Lewkonia P, Bischak DP, *et al*. Using simulation modeling to improve patient flow at an outpatient orthopedic clinic. *Health Care Manag Sci* 2011;14:135–45.
8. Cayirli T, Veral E, Rosen H. Designing appointment scheduling systems for ambulatory care services. *Health Care Manag Sci* 2006;9:47–58.
9. Dexter F. Design of appointment systems for preanesthesia evaluation clinics to minimize patient waiting times: a review of computer simulation and patient survey studies. *Anesth Analg* 1999;89:925–31.
10. Harper PR, Gamlin HM. Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum* 2003;25:207–22.
11. Williams KA, Chambers CG, Dada M, *et al*. Patient punctuality and clinic performance: observations from an academic-based private practice pain centre: a prospective quality improvement study. *BMJ Open* 2014;4:e004679.
12. Gross D, Shortle JF, Thompson JM, *et al*. *Fundaments of queueing theory*. 4th edn. New York: Wiley Publishing, 2008.
13. Hasija S, Pinker E, Shumsky RA. Work expands to fill the time available: capacity estimation and staffing under Parkinson's Law. *Manufacturing Serv Operations Manag* 2010;12:1–18.
14. Williams KA, Chambers CG, Dada M, *et al*. Using process analysis to assess the impact of medical education on the delivery of pain services: a natural experiment. *Anesthesiology* 2012;116:931–9.
15. Williams KA, Chambers CG, Dada M, *et al*. Applying JIT principals to resident education to reduce patient delays: a pilot study in an academic medical center pain clinic. *Pain Med* 2015;16:312–18.
16. Elnahal SM, Moningi S, Wild AT, *et al*. Improving safe patient throughput in a multidisciplinary oncology clinic. *Physician Leadersh J* 2015;2:56–64, 62, 64–5.
17. Thomas S, Glynne-Jones R, Chait I. Is it worth the wait? A survey of patients' satisfaction with an oncology outpatient clinic. *Eur J Cancer Care* 1997;6:50–8.
18. Maiser D. The psychology of waiting lines. In: Czepiel JA, Solomon MR, Surprenant CF, eds. *The service encounter: managing employee/customer interaction in service businesses*. Lexington, MA: Lexington Books, 1985:113–23.
19. Brunk HD. *An introduction to mathematical statistics*. Xerox Corporation, 1975.