HEALTH SYSTEMS
https://doi.org/10.1080/20476965.2023.2215848

THE OPERATIONAL RESEARCH SOCIETY
Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Estimating the workload of a multi-disciplinary care team using patient-level encounter histories

Ekin Koker[a], Hari Balasubramanian[b], Rebecca Castonguay[c], Aliecia Bottali[d] and Aaron Truchil[e]

[a]Independent Data Scientist, formerly PhD student at the University of Massachusetts, Amherst, USA; [b]Mechanical and Industrial Engineering, University of Massachusetts, Amherst, USA; [c]Repair Engineer at General Electric (GE) Renewable Energy, formerly an undergraduate research assistant at the University of Massachusetts, Amherst, USA; [d]Global Sourcing Analyst at Hologic, Inc, formerly an undergraduate research assistant at the University of Massachusetts, Amherst, USA; [e]Director of Strategy and Analytics, Camden Coalition of Healthcare Providers, Camden, NJ, USA

ABSTRACT

Healthcare spending in the United States is concentrated on a small percentage of individuals, with 5% of the population accounting for 50% of annual spending. Many patients among the top 5% of spenders have complex health and social needs. Care coordination interventions, often led by a multidisciplinary team consisting of nurses, community health workers and social workers, are one strategy for addressing the challenges facing such patients. Care teams strive to improve health outcomes by forging strong relationships with clients, visiting them on a regular basis, reconciling medications, arranging primary and speciality care visits, and addressing social needs such as housing instability, unemployment and insurance. In this paper, we propose a simulation algorithm that samples longitudinal patient-level encounter histories to estimate the staffing needs for a multidisciplinary care team. Our numerical results illustrate multiple uses of the algorithm for staffing under stationary and non-stationary patient enrollment rates.

ARTICLE HISTORY
Received 26 April 2021
Accepted 15 May 2023

KEYWORDS
Care coordination; workload estimation; staffing and capacity planning; simulation

Section 1. Background and motivation

Healthcare spending in the United States is disproportionately skewed: just 5% of the population accounts for 50% of annual spending, while just 1% accounts for almost a quarter of annual spending (AHRQ, 2010). Many individuals among the top 1–5% of spenders have complex medical and social needs. According to the National Center for Complex Care, "people with complex health and social needs experience combinations of medical, behavioural health, and social challenges that result in extreme patterns of healthcare utilisation and cost. They repeatedly cycle through multiple healthcare, social service, and other systems but do not derive lasting benefits from those interactions". Such individuals typically have multiple chronic conditions and experience significantly higher than average hospital utilisation rates, including avoidable hospitalisations. The social context – the absence of stable housing or strong support networks, the lack of employment or insurance, the presence of disabilities, and the lack of transportation options, to name a few – further complicates the care delivery process and adds to the individual's vulnerabilities.

Care coordination interventions are one strategy for addressing the challenges facing individuals with complex medical and social needs. Care coordination refers to a concerted effort, often consisting of a multidisciplinary team of nurses, community health workers and social workers, to help improve the health and wellbeing of such patients. While care coordination has largely been employed telephonically, there has been a growing shift to "high-contact", in-person-based interventions. In "high-contact" care coordination efforts – which are the focus of this paper – the care team spends significant time with and on behalf of each patient. For example, members of the care team repeatedly visit the patient out in the community (to observe patient's circumstances beyond what may be visible in a healthcare setting), checking vital signs and discussing symptoms, helping reconcile their medications, arranging primary and speciality care visits, accompanying the patient to those visits, arranging for transportation, procuring medical equipment, etc. Social workers and community health workers in the care team address issues, such as lack of housing, employment, legal services, insurance, mental health and addictions. Interventions can last anywhere between a few weeks to months, and seek to end when patient goals are achieved and patients become self-reliant: their health status has improved, and they are able to go about their daily lives with minimal support from the care team. The hypothesis underlying care coordination interventions is that the care team can help patients (1) improve self-efficacy and well-being by enabling medical and social support, including access to primary and speciality care

CONTACT Hari Balasubramanian hbalasub@admin.umass.edu University of Massachusetts, Amherst, USA

appointments, housing, employment and health insurance; (2) aim to reduce adverse and costly events, such as avoidable emergency visits, medication-related complications, and hospitalisations and increase utilisation of outpatient services and other non-emergent medical care.

This paper deals with aggregate capacity planning and staffing questions in "high-contact" multidisciplinary care coordination teams. While care coordination has been adopted by a range of organisations, including payers, primary care providers, and hospitals, there is a lack of studies on how care teams should be staffed. Staffing a care coordination team that addresses the needs of medically and socially complex patients in a timely manner is challenging for many reasons. First, the number of weeks a patient will be in the care intervention varies significantly from patient to patient and is difficult to predict a priori. Second, a patient needs repeated interactions of uncertain durations with multiple staff types during the length of the intervention (interventions typically last weeks or months). For example, two members of the care team may visit the patient's home each week in the early stages of an intervention; and a home visit can last anywhere from 30 min to 2 h. Third, the number of hours a particular staff type in the care team will need to spend with a patient is uncertain and varies by the week of the intervention the patient is in. For example, registered nurse demands peak in the first 2 weeks of a patient's intervention while SW demands peak in the later stages, once urgent clinical issues have been resolved (Martinez et al., 2019). Thus, varying intervention lengths, recurring visits with uncertain durations for each visit, variability in demand over time via hospitalisations that include peaks and droughts, and multiple staff types who play different roles at different stages all make staffing decisions challenging.

In this study, we focus on an aggregate capacity planning question: How many hours of each care team staff type are needed each week and how does this vary given the number of patients enrolled by the care team each week? Planning for fewer staff hours than what patients demand delays coordination tasks and thereby increases risk of hospitalisation, whereas having more staff hours than necessary causes underutilisation and increases the staffing cost of the intervention. A first step achieving the right balance is the estimation of weekly workload distributions of each staff type as a function of the number of the patients enrolled each week. We demonstrate how granular patient-level encounter data can be used within the framework of simulation to create weekly workload histograms for each staff type. Our data come from the Camden Coalition of Healthcare Providers (hence forth the Camden Coalition), an organisation with significant experience, rich data and national renown in the field of complex care, based in Camden, New Jersey.

The simulation methodology we propose in this paper is necessary due to complexities in the queueing network underlying the care coordination systems such as non-stationary routing and service time parameters. Analytical queueing network models often require Markovian properties which do not apply in care interventions; and while a discrete event simulation of the process can be constructed in theory, it requires significant amount of statistical estimation related to the time-dependent dynamics. Analysis and optimisation in these settings can also be achieved via random field models (as described in some of the papers in our literature review) which describe non-Markovian and non-stationary patient trajectories as well as an "offered load approximation" approach that can be modelled in a mathematical programming framework. However, reliably characterising such patient trajectory models requires a much larger dataset than ours. We have a highly granular dataset; however, it has a small number of patients, which makes it difficult to use these models. We note that small datasets are a common feature of programs that assist patients with complex medical and social needs who typically represent 1–5% of the population.

In our simulation methodology, we instead use the complete longitudinal record of each patient which embeds within it historically observed encounters, their durations, and staff involvement. By randomly sampling the encounter histories of the patients according to the weekly enrolment rates and calculating the superposition of these histories over many replications, we estimate the demand for each staff type in any given week. We demonstrate three different capacity planning uses of the simulation algorithm. First, we estimate the workload histograms for multiple staff types in steady state under a given mean weekly enrolment rate. Second, we show that the methodology can be used to infer joint workload distributions for multiple staff types that are particularly relevant to planning home visits. This is because home visits typically involve different staff types – for example, a licenced practice nurse (LPN) often visits along with a community health worker (CHW) – creating a correlated workload pattern. Finally, we consider the more realistic case where the mean weekly enrolments can change with time.

The rest of the paper is organised as follows. In Section 2, we review the literature around care coordination and staff capacity planning. In Section 3, we explain the Camden core model for care coordination. In Section 4, we summarise the data and give examples of patient-level encounter histories. In Section 5, we conceptualise the intervention as a complex queueing network and present our simulation algorithm. In section 6, we investigate the results of our computational experiments. In section 7, we conclude the paper and map the directions for future research.

Section 2: Literature review

The impacts of care coordination on patients have been studied across numerous types of diseases and patient characteristics. Children are often a focus of care coordination research as they rely heavily on adults in their life for support. A study found that there is a positive association between care coordination and reduced functional disabilities among children with special health care needs. These results were enhanced when services were given in a family-centred medical home, implying that the family is recognised as the primary caregiver and the care coordination team is present to aid and support patients and their families (Litt & McCormick, 2015). Another such study focused on children with medical complexity and sought to analyse the impact of nurse availability and contact – referred to as the nurse dose – on the success of care coordination efforts for children with medical complexity.

Nurses are a key factor in the success of care coordination as they provide a link between patients, their families, and other medical professionals (Cady et al., 2015). Nurses are also crucial to the care coordination context that we study; our model generates staffing estimates for both registered nurses (RNs) and licenced practice nurses (LPNs). Patients also need assistance in sectors beyond healthcare – for example employment, transportation, insurance, housing, and legal services. For this reason, care teams have tended to include community health workers and social workers. As an example, in the intervention described in Powers et al. (2020), the care team consisted of a community health worker, a social worker and a primary care physician. In our case, care teams are led by a licenced practice nurse (LPN) and a community health worker (CHW) and are supplemented by registered nurses (RNs) and social workers (SWs).

Features of care coordination that are most effective in practice have also been discussed in the literature. Brown et al. (2012) noted that successful efforts involve repeated in-person encounters between patients and the care team; medication management; and care team members closely coordinating in person and over phone with a patient's providers. A similar review of care coordination emphasised the importance of patient-centric plans that integrate disease, lifestyle, and behavioural management to increase patient engagement and care effectiveness (Mattke et al., 2015). The team-based care coordination program that we study in this paper has all these features; see Section 3 for a more detailed description on the intervention. Brown et al. (2012) also found that savings can be generated if sufficient funding is provided for care coordination because the reduction in hospitalisation is enough to cover the monthly fees of a care coordination team. For a recent review of the types of interventions and their impact on patient utilisation and cost outcomes, we point the reader to Chang et al. (2023).

Staffing levels are crucial in ensuring that patient needs are met in a timely manner, yet we note that none of these studies in the clinical and health services literature explicitly address the staffing of multidisciplinary teams. One reason for this is that care coordination is an emerging field, and availability of data is limited. Staff scheduling, on the other hand, has been a thoroughly studied concept for decades within healthcare systems, such as emergency departments, hospitals, operating rooms, and physician offices. Queueing theory, newsvendor models and other operations research techniques have been used to tackle scheduling issues (Barz & Rajaram, 2015; Brandenburg et al., 2015). Appointment scheduling and sequencing is especially important in an ambulatory setting where there is a need for emergent care and scheduled procedures (Ahmadi-Javid et al., 2017; Cayirli et al., 2006; Gupta & Denton, 2008).

This paper brings a capacity planning and staffing perspective to the emerging field of care coordination. A recent review paper in Manufacturing & Service Operations Management (MSOM) (Keskinocak & Savva, 2020) highlights "better integrated patient care" is an area of opportunity for future research. They believe "initiatives that aim to better coordinate acute hospital care with preventative and chronic care in the community" could benefit from "data-driven methods" which is exactly what our paper aims to do.

Papers from OR&OM (Operations Research and Operations Management) literature that are closest to ours are Campello et al. (2017), Chow et al. (2011), Deglise-Hawkinson et al. (2020), Helm and Van Oyen (2014), Hilton et al. (2018), Howells et al. (2022) and Rossi and Balasubramanian (2018). In what follows, we review each paper and its relationship to our study.

A key feature tackled in our paper is repeated interactions between patients and care team staff. Similarly, Campello et al. (2017) model the interactions of customers with "case managers" using queueing theory. They define case managers as servers who are assigned multiple customers and have repeat interactions with those customers. They give examples of ED doctors, customer service representatives using online chat, and social workers. Members of the care team in our study operate in a manner similar to case managers. The difference is that case managers act independently whereas care coordination teams, which include nurses, community health workers and social workers among others, work together to help a single patient. Campello et al. (2017) assume homogeneity of customers and servers and assign Markovian properties to arrival and service rates. Our study assumes both deterministic and Poisson arrivals that can be stationary or non-stationary; for

service rates we sample patient-care team encounter histories and use the service time realisations in these encounter histories.

Hilton et al. (2018) models paediatric asthma patients who also have repeated interactions with the healthcare system following an ED visit or hospitalisation after an asthma attack. They use Markov renewal processes to summarise time-ordered events with varying time intervals between events (e.g., hospitalisations, ED visits, physician office visits) and use model-based clustering to create patient profiles and visualise them using network analysis. It is a paper that describes key patterns, similar to our earlier work in Martinez et al. (2019) but does not offer staffing implications as we try to do here.

Howells et al. (2022) models an adult psychology clinic in U.K'.s National Health Service using a discrete event simulation. Care coordinator roles are an important part of psychology services, and this paper identifies the bottlenecks and suggests different staffing scenarios to improve accessibility of patients to mental health services. Their results indicate that having some of the therapy staff take on the role of dedicated care coordinators could improve the outcomes of the clinic.

Chow et al. (2011) uses Mixed Integer Programming and an uncapacitated Monte Carlo simulation to optimise surgical scheduling to reduce ward congestion. They sample patient trajectories from a database and also try to predict staffing/utilisation levels using MIP based optimisation and derive guidelines for scheduling from the optimised solutions. Their patient flow structure is standardised, as patients go through pre-determined wards, typical in a hospital setting, as opposed to the highly variable patient trajectories through multiple staff types that is a feature of our care coordination setting. Chow et al. (2011) use of *trace-driven simulation* – i.e., they directly sample patient – level timestamps from historical data. Chow et al. (2011) point out that trace-driven simulations have a drawback in that they can only reproduce historical observations; however, they also note that "this method can preserve correlation patterns between patient type, length of stay, and patient path for each patient". Maintaining such patient-specific correlations and the non-stationary dynamics of an intervention, which are difficult to characterise in an analytical patient trajectory model, is precisely why we also chose this method.

We note further that practical operations work in hospitals focuses on ensuring enough staff hours are available to meet the historical demand. The assumption is that having staff present for certain hours on a certain day is sufficient for treatment needs. The need for multiple staff types that jointly assist the same patient over a period of time is often not considered; and in fact, data on time spent by nurses and other non-physician staff (such as medical technicians and social workers) is not always available in hospital settings. Fortunately, in our care coordination setting, data on the timing and duration of patient-care team encounters is quite detailed. This allows to capture of how multiple staff types are involved during the course of a patient's intervention. For example, community health workers (CHWs) and licenced practice nurses (LPNs) regularly conduct home visits together. The sampling of patient specific encounter histories allows us to capture these correlated work patterns and the joint distribution of hours needed.

Helm and Van Oyen (2014) use a random field model and stochastic mixed integer programming to optimise the admission scheduling and control problem for an entire hospital, with the goal of stabilising the hospital census. A key aspect of their modelling framework is a characterisation of a patient's trajectory from arrival to the different inpatient wards in the hospital. Specifically, patient-level data is converted into a probability distribution that can change over time: it captures the probability that a patient of a certain type (e.g., a cardiology patient, or an elective surgery patient) would require a bed in a particular ward on day d since arrival. This analytical model of a patient's evolving trajectory in the hospital is both non-Markovian and non-stationary and is relevant to our study. However, parameterising such a model in the care coordination setting is a challenge. In large hospitals, where thousands of patients are admitted in inpatient wards each year, patient trajectories are easier to estimate when compared to care coordination interventions where only a small number of medically complex patients are enrolled.

Deglise-Hawkinson et al. (2020) model clinical research operations mathematically using a method they call CAPTAIN (CApacity Planning Tool And INformatics). A clinical research patient participating in a trial goes through a specific protocol that involves repeated visits to a research unit. Repeated patient visits are also a feature of care coordination. CAPTAIN considers the clinical research trials planning and scheduling problem from multiple perspectives. It determines which new trial(s) to take on in a heterogeneous portfolio (and which to refuse) while considering how the required visits (determined by the specific protocol) will transpire over time and also ensuring that physical and nurse resources are not exceeded. The CAPTAIN framework also captures the Time to First Available Visit (TFAV), the earliest available day the patient's first visit can be scheduled. In addition, the model allows for nurses with different skill sets and other aspects, such as procedure rooms.

The difference with our system is that a clinical research trial is much more prescribed and a lot less variable compared to care coordination. In particular,

the interval between visits is pre-specified by the protocol with some flexibility around the precise day (for example, the next visit needs to happen between 26 and 30 days from the current one, with the precise day having a uniform probability of being chosen by the patient). In care-coordination, the interval between visits is not pre-specified, it must be estimated from historical data. These intervals are stochastic and can be short (e.g., encounters happen on successive days) or long (no encounters occur for weeks). Longer intervals are more likely to occur in the middle or later stages of the intervention, when the patient's health has stabilised. Another key difference is that the workload induced by a visit in Deglise-Hawkinson et al. (2020) is deterministic, while the patient-care team encounter duration is stochastic and depends on the encounter type (for instance, home visits may require up to 2 h while phone calls may only need a few minutes). Thus, there are two levels of stochasticity that need to be accounted for in our model: one at the level of intervals between visits (measured in terms of number of days), and the other related to the precise duration of the encounter. This increases the estimation burden of an analytical patient trajectory model significantly.

In summary, Helm and Van Oyen (2014) and Deglise-Hawkinson et al. (2020) use exact optimisation approaches in the hospital and clinical research trial settings, respectively. Embedded within their optimisation framework are analytical models of a patient's flow/visits through time and the calculation of the offered workload (i.e., the workload that would be induced in a system without capacity limits). In contrast, we directly sample patient-level encounter histories in a trace-driven simulation and calculate the offered workload realisations for multiple staff types to facilitate aggregate capacity planning. Thus, our approach is a heuristic one designed for a dataset with a small number of patients, but which nevertheless has a high degree of granularity to model a complex sequence of non-stationary longitudinal encounters. We view our aggregate capacity planning model as the first step in the development of more sophisticated approaches.

Finally, Rossi and Balasubramanian (2018) quantifies the workload of a primary care physician (PCP) using longitudinal event histories assembled from the from Medical Expenditure Panel Survey (MEPS). The event histories concern patient visits to primary and speciality care providers as well as emergency department and hospital stays. By randomly sampling event histories for a nationally representative panel of patient, the paper estimates the distribution of two types of workload associated for a primary care physician: weekly face-to-face office visits; the number of weekly non-PCP events, an indirect proxy for the coordination workload for the PCP. Our paper has a similar sampling methodology that uses patient event histories. The main difference is that although both papers try to estimate the demand from a panel of patients, a PCP panel is static in Rossi and Balasubramanian (2018), whereas panel of a care coordination team is dynamic. In other words, in care coordination programs patients are enrolled and eventually complete the program in a few weeks or months, while a primary care panel remains largely static as the PCP builds a long-term relationship (typically many years) with her patients. Thus, care coordination is a much shorter-duration, higher-engagement effort with an ever-changing mix of patients compared long-term, low-engagement effort of a PCP office. A second important difference is that while Rossi and Balasubramanian (2018) model multiple event types, the duration of the event types are assumed since data is not available. In contrast, the Camden Coalition data contain event types as well as durations. Finally, Rossi and Balasubramanian (2018) is concerned with the workload of a single provider (the PCP), while our paper is models the independent as well as joint distribution of the workloads of multiple staff types in the care team.

In summary, interventions for patients with complex medical and social needs are an emerging area without a current knowledge base in the realm of staffing and capacity planning. We create weekly workload histograms for different staff types for a given enrolment rate by sampling the complete history of care team interactions with a patient. The sampling algorithm allows us to efficiently use a large number of details (i.e. high dimensional data) for a small number of patients – a feature common to all complex care intervention datasets – without needing to parameterise a complex patient trajectory model in a queueing network. We demonstrate three different capacity planning applications of the algorithm, including the case where arrival rates can change over time. Thus, our algorithm considers non-stationary dynamics from two different perspectives: non-stationary/time-dependent dynamics of the intervention, which are included in the longitudinal encounter histories of the patients; and non-stationary/time-dependent weekly enrolments (arrival patterns).

Section 3: The core model

We first describe further details of the care coordination intervention at the Camden Coalition. The Coalition's 'Core Model' care coordination program (CM) works with medically and socially complex patients who frequently utilise the hospitals in Camden, a medium-sized city on the East Coast and

one of the poorest and under-resourced cities in the country. The patients selected for the community-based clinical and social coordination program are supported by teams consisting of non-physician members for 30 to 120 days following a period of repeated hospital utilisation. Individualised care plans that are co-created by patients and care team members shape the duration and course of the intervention and are aimed at building durable connections between patients and the medical and social community resources. The importance of "authentic healing relationships" to run an effective care coordination programme is highlighted by the patients in CM.

Patients are considered for CM if they are currently hospitalised, have experienced two or more inpatient admissions within a six-month period, and have 2 or more chronic conditions. To qualify, they must also have 2 or more additional barriers, including but not limited to: polypharmacy (5+ medication), lack of social support, housing instability, active drug use, physical disabilities (e.g. hearing or vision impairment), difficulty accessing serves (e.g., language barrier, limited mobility, lack of transportation), and significant mental health conditions.

Some types of hospital admissions do not satisfy the eligibility criteria. Some examples include admissions for oncology treatment, surgery, acute trauma (e.g., motor vehicle accident), chronic illnesses with limited treatment options (e.g., multiple sclerosis), and mental health treatment only. Moreover, patients who do not have the mental capacity to consent to the program, who are permanent residents of a facility, or who are over the age of 80 are excluded.

CM starts with a patient-centred care planning process in the hospital. During this, the needs of the patient are organised using a taxonomy of 16 care planning domains, such as housing, addiction, and legal. After patient is discharged from the hospital, CM team plans to meet patient in their home within five days of discharge to continue the process and perform a medication review by a nurse before a physician reconciliates their medication. Then, team plans to reconnect the patient to primary care within 7 days of discharge and accompany them to their appointment. Afterwards, the team establish contact with the patient every week as they support the clinical and social goals of the patient by coordinating their care. When team determine that a patient has reached their goals and developed sustainable connections to resources that can support them medically and socially, the patient graduates from the program. Otherwise, the enrolment is deemed incomplete if a patient cannot be located, is no longer interested in receiving CM services, moves outside of the program's geographic boundaries, is incarcerated for an extended period, enters a long-term care facility, or is deceased. In some situations, a patient who did not finish the intervention may be re-enrolled if they are readmitted to the hospital.

The CM staffing model evolved over time, but operated primarily through assigning each patient a two-person team consisting of a Licensed Practical Nurse (LPN) and a Community Health Worker (CHW). These two-person teams were supplemented by a Registered Nurse (RN) who spanned across teams and helped compliment the LPN for certain patient events, such as care planning and initial home visits, that were generally clustered earlier in the intervention. In addition to the RN, the care teams were also supported by a shared Social Work (SW) team. The RNs and LPNs focus on clinical coordination, such as escorting the patient to primary care or speciality appointments. The CHWs are responsible for social coordination, for instance, helping patients connect with community resources, engaging with family members, and sub-acute rehabilitation facility rounds. Social workers and a clinical psychologist are consulted as the need for their expertise arise, for example, events that might require behavioural healthcare or advanced social coordination, e.g., housing arrangements. Americorps Community HealthCorps (which is now defunct) volunteers supplemented the CM team during the study period, and they were responsible for work requiring less complex social coordination.

Section 4: Data and examples of patient-level encounter histories

We first provide a close look at the type of data used in our study and which forms the basis for the simulation methodology in Section 5. We use patient-care team encounter records for 531 patients enrolled into the Camden Coalition programme for over 164 weeks beginning in 2012. These records described 24,249 h of staff effort. Care coordination encounters conducted with or on behalf of patients were recorded by care team members in the field via tablet computer. This research involves retrospective analysis of de-identified patient data. It was approved by the Institutional Review Board (IRB) at the corresponding author's university. The protocol number is IRB: #198 2010–0722; a waiver of informed consent was obtained as part of the protocol.

This section provides examples of how the Coalition's dataset is used to recreate patient-level encounter histories. Consider three patients' timeline of events leading up to the first primary care physician visit in Figure 1. Each shape and colour represents a specific event and staff type. For example, the blue triangle in week 23 of patient 168 shows that a social worker had an interaction with this patient in the hospital. The stacked symbols represent multiple interactions that take place over the course of

a single day. If multiples of the same shape are stacked within a single day, it indicates that more than one staff type was present during the patient event: encounters involving more than one staff are quite common. For example, the three different coloured circle on day 3 of patient 177 indicates that three separate staff types attempted an enrolment visit together.

In addition, the dataset also contains the amount of time reported by the staff for a particular encounter. For example, a registered nurse (RN) and community health worker (CHW) might spend 1.5 h together with the patient on a home visit. Or, a social worker might spend 2 min calling an agency on behalf of the patient. Thus, the dataset includes not only the type of encounter and staff involved but also the time spent during the encounter by the staff. All encounters on behalf of a patient are recorded, whether patient is present during an encounter or not. In all, 27% of 24,249 h of staff effort happened without the patient being present while 3% occurred with the patient over the telephone. Activities such as administrative meetings and trainings that were not liked to specific patients are not included in the dataset.

This detailed history allows us to build encounter histories for each patient for the entire duration of the intervention as well as the cumulative progression of care team hours for each patient. As an example, the cumulative progression for two sample patients is shown in Figure 2. The x-axis shows the number of days passed since their enrolment in the program (negative days specify pre-enrolment activity) and the y-axis shows how many hours the staff members spent with the patient in total cumulatively. The slope of the trajectory represents the intensity of the staff effort: steeper sections indicate that the staff members are spending more time with the patient while flatter portions indicate minimal care team effort. In the above example, both patients start with a steep trajectory but the slope for the Individual 262 stays higher until day 300 and then tapers off at the end whereas for the Individual 229, the slope decreases (in a relative sense) between days 200 and 300 and then increases sharply again before conclusion of the intervention. Note that the graph goes into the negatives due to the effort spent by the care team prior to the enrolment – generally the care team meets the patient at the hospital bedside in an effort to enlist them in the intervention.

As noted earlier, care coordination intervention durations and effort can vary significantly from patient to patient. Accurately predicting the length of intervention and care team hours at the beginning of the intervention is challenging since each patient has a unique mix of clinical and social needs. The average duration of an intervention in the Coalition's dataset is 15.43 weeks, however, the 10^{th} and 90 percentiles are 4 and 31 weeks, respectively. Similarly, the average number of hours spent by all members of the care team on a patient is 46.1 h, while the 10^{th} and 90^{th} percentiles are 6.8 h and 101.77 h, respectively.

More detailed analysis and categorisation of patient trajectories, and descriptive analysis of staff effort and encounter types is provided in [Martinez et al. 2019]. The following points summarise the most findings relevant to this study:

(1) Staff effort is highly front loaded: One fifth of post-enrolment staff effort was delivered within 12 days of enrolment (i.e., beginning of care coordination intervention), and two fifths were delivered within 33 days, while the next two fifths were delivered from day 34–117.

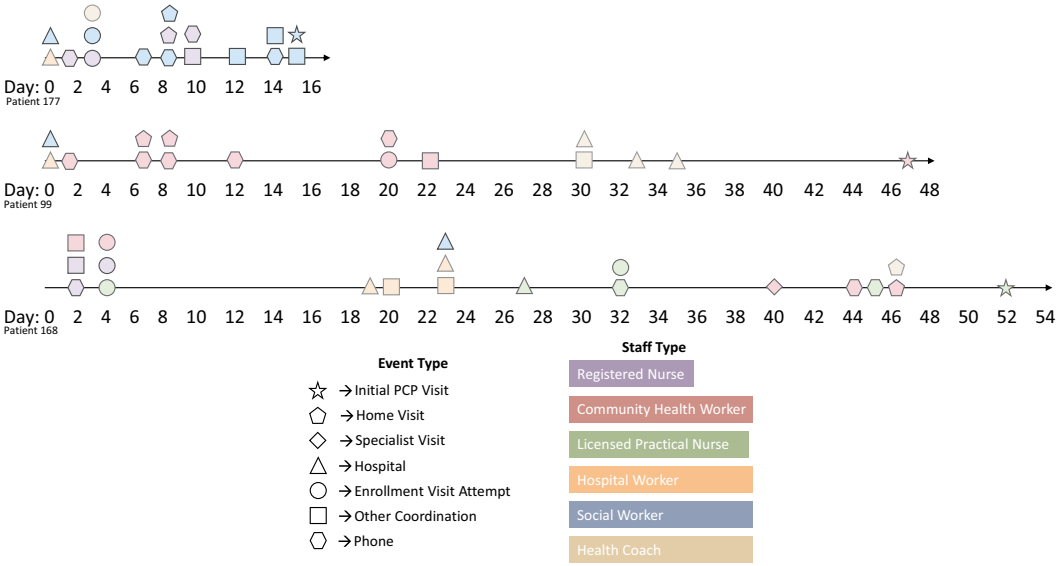

Figure 1. Timeline of events for three patients until their first primary care visit. (Color online).

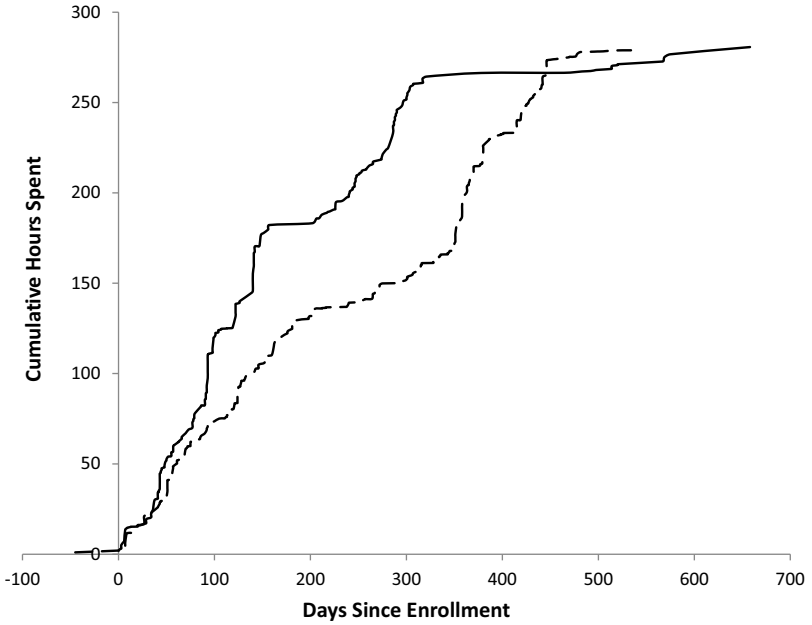

Figure 2. Cumulative care coordination hours by day of intervention for two sample patients.

(2) Staff effort is not uniformly distributed across staff types: registered nurses (RNs), community health workers, social workers, licenced practice nurses, and health coaches contributed to 4.2%, 27%, 16.1%, 24.5%, and 25.2%, respectively.
(3) Different staff types have different effort profiles. As an example, figure 3 shows the average number of RN and SW hours needed by week of intervention for the 526 patients in the dataset. The early weeks of the intervention are the busiest for both RNs and SWs, however RN effort drops more steeply. This is because RN effort focuses on clinical needs that are vital immediately after the intervention to avoid readmissions, while SW efforts on social aspects continue once the immediate clinical needs have been fulfilled.

Table 1 summarises the staff effort in hours for different encounter types. The table demonstrates that staff types vary significantly in terms of the number of hours they spend with patients. Community health workers have the highest effort and registered nurses the lowest. The table also demonstrates how certain encounters require more time from certain staff types compared to others. For example, PCP Visit and Specialist Visit indicate encounters where

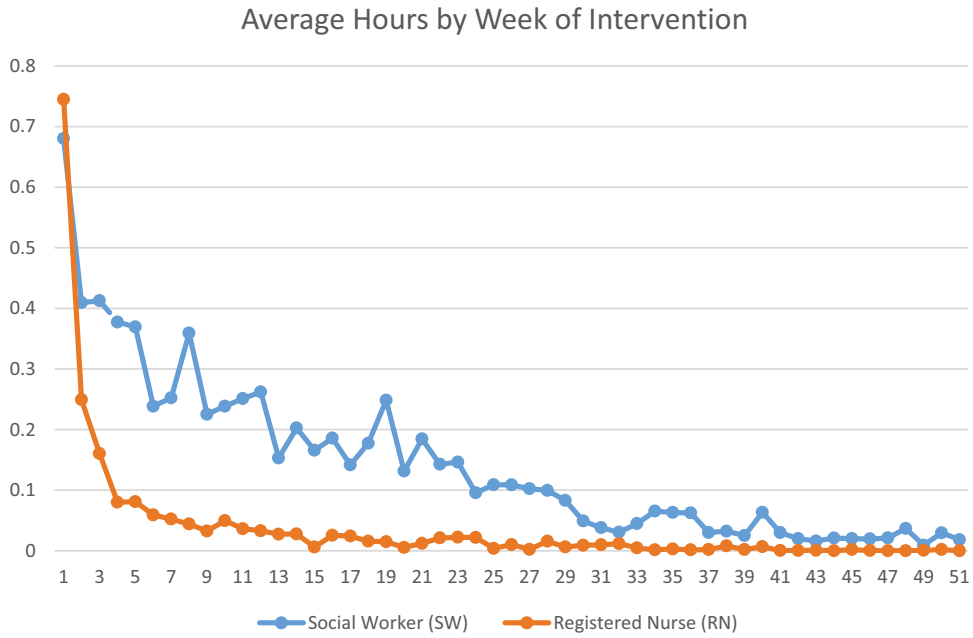

Figure 3. SW and RN average effort profiles by day of intervention. (Color Online).

Table 1. Skill matrix (encounter type vs. hours spent per staff type). Staff types listed: CHW (community health worker); HC (health coach); LPN (licenced licenced practical nurse); RN (registered nurse); and SW (social worker).

Row Labels	CHW	HC	LPN	RN	SW	Grand Total
Home Visit	2520.51	1697.57	2130.5	265.75	759.08	7373.41
Community/Nursing Home/Other Facility	651.25	1108.28	546.01	65.67	1244.64	3615.85
Enrollment Visit	473.5	308.75	608.75	315.83	52.75	1759.58
Clinical Coordination	636.87	422.92	423.16	56.31	43.3	1582.56
Social Coordination	94.01	408.36	51.94	12.03	957.99	1524.33
Specialist Visit	391.75	294.25	417.92	19.5	45.58	1169
Hospital	168.33	358.08	228.09	55.75	313.59	1123.84
Phone	266.57	353.15	211.2	34.12	134.57	999.61
PCP Visit	476.5	317.58	621.92	65.75	85.34	1567.09
Other	867.74	832.97	702.05	136.2	270.16	2809.12
Grand Total	**6547.03**	**6101.91**	**5941.54**	**1026.9**	**3907**	**23524.39**

a care team member accompanies a patient to the primary care and speciality care office. We see that social workers have the lowest involvement in PCP and Speciality visits but have the highest hours when it comes to meeting patients in alternative settings in the community, for instance nursing homes.

Section 5: Methodology

5.1. Multidisciplinary care intervention as a non-stationary queueing network

The care intervention process for patients with complex medical and social needs can be conceptualised as a non-stationary queueing network, as shown in Figure 4. In the network, each care team staff type – registered nurse or RN; community health worker, or CHW; social worker, or SW; and licenced practical nurse, or LPN – serves as a node that is numbered accordingly. We use these four staff types as an example while the actual problem also includes other staff types, such as health coach and clinical psychologists. Patients are enrolled into the intervention at a rate of λ per unit time. Once enrolled, a patient's first interaction could be with either of any of the care team staff types; these node-specific arrival rates are denoted by λ_j where j = 1,2,3 and 4 in the figure. Note that $\lambda = \sum \lambda_j$. In Figure 4, arrivals to each of the staff type nodes are represented by the dashed lines. Once the patient *i* is enrolled, she has recurring encounters with multiple staff types in the care team for the length of the intervention, similar to the three examples of event progressions shown in Figure 1.

Let t denote a moment of time at which we observe the queueing system, and let W_t denote the set of patients who are active in the intervention at t. Denote by $\tau_{i,t}$ the number of days or weeks patient i has been in the intervention. We use t and $\tau_{i,t}$ because each active patient's needs vary depending on the stage of the intervention. At each node j visited by the patient i at time t, the relevant care team member spends $Z_j(\tau_{i,t})$ hours with the patient i. The patient then moves from node j to node k (transitions to the same node are also allowed) with probability $p_{j,k}(\tau_{i,t})$; furthermore, this transition takes $G_{j,k}(\tau_{i,t})$ days, which reflects the intervals between successive encounters.

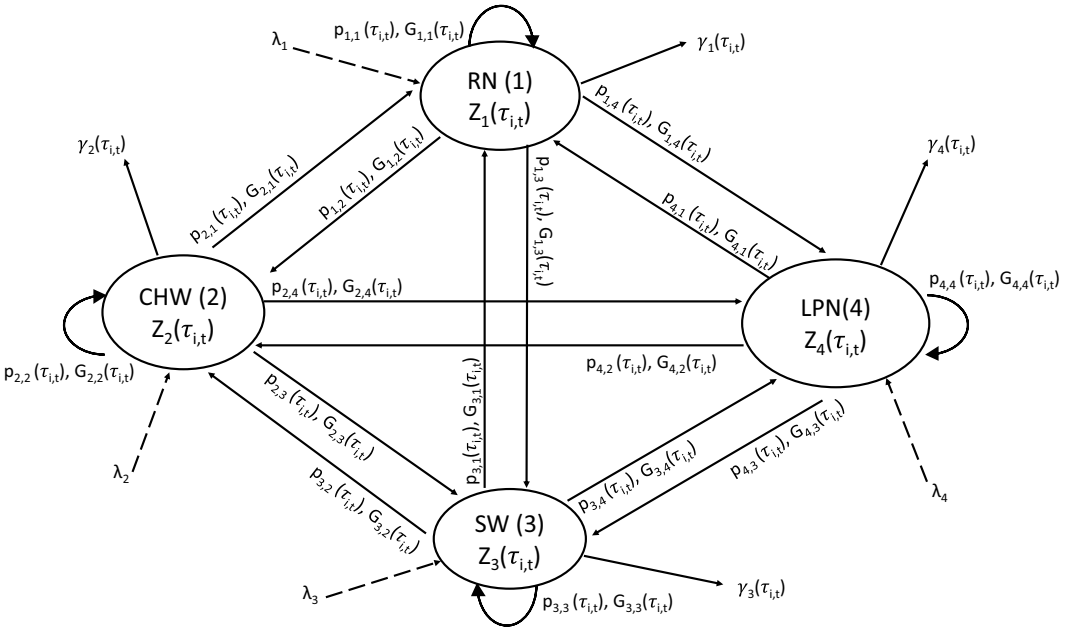

Figure 4. Care team intervention conceptualised conceptualised as a non-stationary queueing network.

Figure 1 provides examples of such transitions and intervals between transitions for three patients until their first primary care visit. From any node, the patient can also exit from the system with probability $\gamma_{j,k}(\tau_{i,t})$, where $\gamma_{j,k}(\tau_{i,t}) = 1 - \sum_k p_{j,k}(\tau_{i,t})$. Exiting the system is equivalent to completing the intervention.

As discussed earlier and illustrated by Figure 3, earlier stages require greater care team time and frequency of interactions, compared to later stages. Furthermore, certain staff types such as registered nurses are needed earlier in an intervention, while social workers are needed later. Thus, the queueing network described here has non-stationary dynamics even if the mean enrolment rates are unchanging. Therefore, analytical queueing models such as Jackson queuing networks which assume stationary and memoryless transition probabilities cannot be used. Memory in this case refers to knowledge of the stage of the patient's intervention, captured by $\tau_{i,t}$.

We are interested in capacity planning for each staff type in this queueing network. Specifically, for an enrolment rate of λ what is the demand/workload distribution, i.e., the number staff hours needed on a daily or weekly basis, at each node? Once such a distribution is estimated, we can determine the capacity levels at which each staff type should work to ensure that patient needs are met on a timely basis and staff are not too underutilised.

One approach to estimating the demand distribution at each node is to use discrete event simulation using off-the-shelf software. However, building such simulation requires a significant amount of input modelling, specifically estimating the time-dependent parameters, such as $Z_j(\tau_{i,t})$, $p_{j,k}(\tau_{i,t})$ and $G_{j,k}(\tau_{i,t})$. The variation in registered nurse and social worker hours by stage of intervention seen in Figure 3 suggests that service time and routing parameters would need to be estimated for each week the patient is in the intervention. Such estimation requires a non-trivial amount of computational effort, formulating and optimising maximum likelihood functions for $Z_j(\tau_{i,t})$, $p_{j,k}(\tau_{i,t})$ and $G_{j,k}(\tau_{i,t})$, using data observed for each stage of the intervention (possibly for each week based on Figure 3). While $Z_j(\tau_{i,t})$ is relatively easier to estimate, jointly estimating $p_{j,k}(\tau_{i,t})$ and $G_{j,k}(\tau_{i,t})$ – the routing matrix and the intervals between successive visits – poses a significant challenge. Essentially $p_{j,k}(\tau_{i,t})$ and $G_{j,k}(\tau_{i,t})$ together represent a patient's non-stationary, probabilistic trajectory through the care intervention process, which, in the terminology of stochastic processes, can be called a *non-stationary renewal process* involving transitions between multiple states/events. Parameterizing such renewal processes requires datasets with a large number of patient trajectories while the number of patients in care intervention programs is typically small.

The queueing network conceptualisation discussed above also does not include an important feature of complex care interventions: multiple staff types often need to be present for certain patient encounters. Home visits, for example, often require two different staff types – registered nurse and social worker; or licenced practical nurse and community health worker – to visit the patient's home together. Thus, the service time $Z_j(\tau_{i,t})$ is often jointly shared across two different staff types.

Therefore, rather than using traditional queueing frameworks, which does not accommodate non-stationary dynamics that include memory, a discrete event simulation which requires significant parameter estimation, or more recent analytical approaches towards patient trajectory characterisation (Deglise-Hawkinson et al., 2020; Helm & Van Oyen, 2014) we instead use an approach that samples the complete longitudinal event history of individual patients. The complete longitudinal record of a historical patient flow data set embeds within it all of the available information for modelling the non-stationary/time-dependent dynamics: the variation of routing and service time parameters are automatically captured without the explicit estimation of parameters. Furthermore, since at any time period (a day or a week), patients can be in many different stages, we can use the principle of aggregation/superposition by summing the demands across the sampled patients who are still active in the intervention. Our method focuses on a practical modelling approach that can capture the significant complexity of the patient flow processes in a manner that is useful for practitioners in healthcare improvement. In the next section, we describe the intuition as well as the details of our sampling-based simulation algorithm.

5.2. A Patient trajectory-based simulation algorithm

We capture a patient's encounter history with the following notation. Denote by $Z^e_{i,w,s}$ the number of hours spent by staff type s on patient i in encounter type e in week w of the intervention. Here, w can range from enrolment week $w = 1$to week$w = \mu_i$, when the intervention completes. Without loss of generality, we have chosen a granularity of a week to allow for some level of aggregation, however, the data allow for a granularity of a day. Let $E_{i,w,s}$ denote the set of all encounters involving patient i and staff type s in week w of the intervention. Then, the total hours spent by staff type s on patient i's

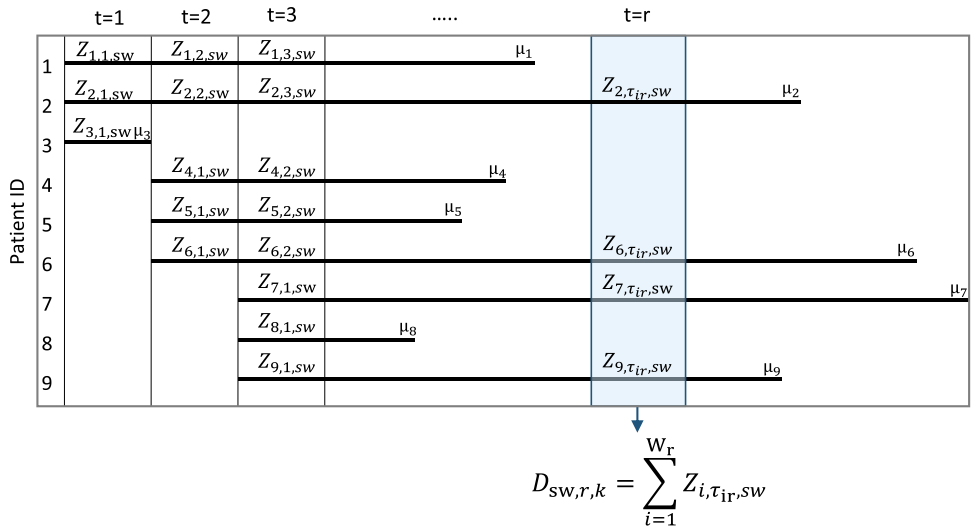

Figure 5. Visual illustration of how the simulation algorithm works for the case where exactly $n = 3$ patients are enrolled every week. Calculation of the demand for staff type sw in week r of replication k is shown.

care when the patient is in week w of the intervention is given by:

$$Z_{i,w,s} = \sum_{e \in E_{i,w,s}} Z^e_{i,w,s}$$

We describe next how this aggregated encounter history can be used in the simulation to create workload histograms for each staff type, as a function of the weekly enrolment rate. These workload histograms show the distribution of the required staff effort and can be used by the team to determine a weekly capacity (number of hours employed for each staff type) as described in the next section. Figure 5 provides a visual overview of the simulation algorithm when exactly $n = 3$ patients are enrolled each week. The algorithm randomly samples n unique patients without replacement (i.e., patients who have not been sampled before) each week from the dataset. Each patient is thus assigned a starting week, and stays in the model for μ_i weeks. In any week therefore, the active patients include those who were enrolled in prior weeks and who are yet to graduate and the newly enrolled patients. This sampling process is repeated until the system reaches steady state and the number of active patients stabilises. The sum of the hours requested of a staff member by patients who are active in any week can be calculated after steady state is reached. This enables estimating weekly workload distribution unique to each staff type. We can think of this simulation as the *method of superpositions of patient encounter histories* to obtain demand estimates. By repeating the simulation algorithm multiple times, we can collect many weekly realisations since the superpositions will involve different combinations of patient encounter histories. In what follows, we provide a formal description of the algorithm, starting with the notation.

Notation

s: index for staff type, $s \in$ {rn, lpn, chw, hc, sw}
t: index for indicating week in simulation
i: index for patients
M: set of indices for patients for whom historical event data exists. For our study, |M| = 526.
n: variable in simulation algorithm that stores number of patients enrolled in each week
μ_i : number of weeks that patient i is active in the intervention
a_t: set of indices of patients newly enrolled (arriving) in week t
l_t: set of indices of patients leaving the programme in week t
W_t: set of all patients that are currently active in week t
$\tau_{i,t}$: the number of weeks a patient $i \in W_t$ has been in the intervention at time period t. For example, if a patient i joined the intervention 3-weeks prior to the current week t, then $\tau_{i,t} = 3$.
K: total replications of the simulation, indexed by k
U_k: set of indices of all patients not sampled prior to week t in replication k of the simulation. $U_k \subseteq M$.
$Z_{i,\tau_{i,t},s}$: time (in hours) needed of staff type s by patient i in week $\tau_{i,t}$ of patient's i's intervention.
$D_{s,t,k}$: total demand (hours) for staff type s in week t of replication k

Initializations for first week of simulation for replication k

Set $t = 1$
Initialize $U_k = M$ *[In the beginning all patients for whom we have data can be sampled.]*
Determine n, the number of patients enrolled in week 1. *[n is either constant or sampled from a Poisson distribution with a predetermined mean.]*
Randomly sample n unique patient indices from U_k. Add the n patient indices to the set a_1.

Initialize $W_1 = a_1$, and $l_1 = \emptyset$ *[Patients active in week 1 are those who were just enrolled. No patients left the programme in week 1, hence this set is empty.]*
Initialize for all $i \in a_1$ $\tau_{i,1} = 1$ *[All newly enrolled patients are in week 1 of their intervention.]*
Update $U_k = U_k \backslash a_1$ *[Patients' indices sampled in week 1 cannot be sampled again]*

Simulation algorithm for replication k

Step 1: Set $t=t+1$ *[Advance the week of the simulation.]*
Step 2: Determine n, the number of patients newly enrolled in week t. [*n is either constant or sampled from a Poisson distribution with a predetermined mean.]*
Step 3: If $|U_k| < n$, go to Step 8 *[End simulation if n unique patients cannot be sampled.]*
Else
Step 4: Randomly sample n unique patient indices from U_k. Add the n patient indices to the set a_t
Step 5: Update the following sets and variables:
$W_t = (W_{t-1} \backslash\ l_{t-1}) \cup a_t$ *[Patients who left the programme in week t-1 (denoted by l_{t-1}) are removed from the set of active patients and patients newly enrolled in week t are added to the set of active patients. Note that \ indicates set subtraction.]*
For all $i \in a_t$, $\tau_{i,t} = 1$ *[Patients who joined in simulation week t are in week 1 of their intervention.]*
For all $i \in (W_t \backslash a_t)$, $\tau_{i,t} = \tau_{i,t-1} + 1$ *[Each patient i who did not newly enrol and is currently active enters the next week of their intervention. In other words, the updated $\tau_{i,t}$ value captures which week of the intervention they are currently in.]*
$U_k = U_k \setminus a_t$ *[Patients' indices sampled in week t cannot be sampled again in replication k.]*
Step 6: Calculate demand or workload for each staff type s in week t by summing all the demand from patient i active in set W_t that are in week $\tau_{i,t}$ of their intervention:
$D_{s,t,k} = \sum_{i \in W_t} Z_{i,\tau_{i,t},s}$

[Demand for staff type s is the sum of the time needed by each patient active in week t of simulation. The time each patient needs from each staff type in week t depends on $\tau_{i,t}$.]
Step 7: Return to Step 1
Step 8: End simulation

5.3. Estimating the steady state demand distribution

Since the simulation starts with the system empty and idle (i.e., at $t = 1$ W_t is empty before patients are sampled), there is a warm-up period before steady state is reached. We now illustrate how to calculate the steady state weekly demand distribution for each staff type. If μ is the average intervention duration (average time from enrolment to graduation), then μ can be estimated by averaging the intervention durations of each patient in the dataset:

$$\mu = \sum_{i \in M} \mu_i / |M|$$

Let λ be the average number of patients enrolled each week.

The mean steady-state number of patients active during the intervention, L, is therefore given by:

$$L = \lambda * \mu$$

This is a straightforward application of Little's Law (Little, 1961) which states that the long-term average number of customers in a stationary system is equal to the long-term average effective arrival rate multiplied by the average time that a customer spends in the system. As new customers arrive and existing customers leave in each period, a certain number of individuals remain active in the system; Little Law's estimates the mean number of active customers. Note that L can be calculated before the simulation begins and therefore it can be used to exclude weeks in the warm-up period and determine when steady state has been reached. More precisely, in any replication k when $E[|W_t|] \approx L$ (i.e., the average number of patients active in simulation week t reaches L) the simulation has reached a steady state.

Let T be the total number of weeks that simulation runs in replication k and b indicate the week in which steady state is reached. Therefore, the steady state demand distribution for staff type is s is given by $D_{s,t,k}$ for all weeks $t = b$ to $t = T$ and for all replications $k = 1$ to $k = K$. For the Camden Coalition's dataset, $\mu =$ 13.3 weeks. For a mean arrival rate of $\lambda = 3$:

$L = 3 \times 13.3 = 39.9$

Therefore, the steady state is achieved when the average number of patients active in the intervention stabilises at approximately 40 patients. Once the number of active patients stabilises, we can estimate the steady state workload distribution for each of the staff types. Figure 6 shows average weekly demand values in hours (based on 5000 simulation replications) for each staff member when exactly 3 patients are enrolled each week. The simulation runs for 175 weeks since beyond this time it is not possible to sample 3 unique patients from the total of 526. The number of patients stabilises at around from week 40 onwards and so do the weekly hours for each of the staff types. Thus, by discarding observations prior to week 40, and by using the observations from weeks 40 to week 175 we can estimate the distribution of demand for each staff type.

For a constant arrival rate of $\lambda = 6$ the steady state is still achieved from week 40 onwards, however the weekly demand hours for each staff type

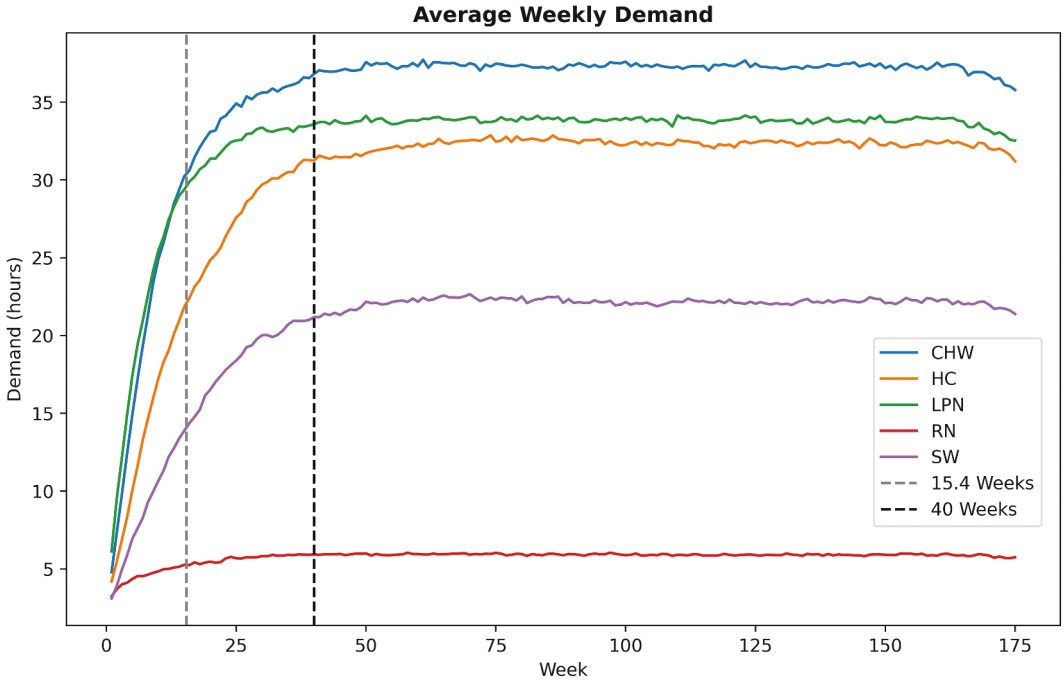

Figure 6. Average weekly demand for each staff type (based on 5000 replications) by week of simulation. (Color Online).

are roughly doubled. In this case, each replication of the simulation runs for 87 weeks in total since we cannot sample 6 unique patients from the dataset from week 88 onwards (by week 87, $87 \times 6 = 522$ patients are sampled and only 3 patients are left). Thus, the steady state distribution for each of the staff types is calculated from week 40 to week 87. This still gives us 47 observations in each replication and if we repeat the simulation 100 times, we get 4700 weekly observations from which to estimate the weekly workload distributions for each staff type.

5.4. Underlying assumptions and limitations

We now address some underlying assumptions and limitations in our simulation methodology.

(1) Biases in the Data: The encounter histories are based on actual realisations, so they may contain delays experienced by patients. In other words, a particular home visit for a patient should have occurred on week 5 of the intervention, but instead occurred in week 6. For some other patients, encounters might have higher tolerance so that while an encounter occurred in a certain week, it could have safely been moved to the following week. To test if such factors might impact our histograms, we conducted simulations in which the encounter dates were "perturbed" with a small probability – that is moved to a week prior or a week later. We found that the superpositions of the perturbed event sequences did not change the workload histograms. This is because in any week, different patient combinations are involved in the superpositions leading to an averaging effect that produces histograms similar to the unperturbed case.

(2) Patient Wait Times: We have not considered patient waiting times in this study. This is because timeliness is of the essence when it comes to patients with complex medical and social needs who have been recently discharged from a hospital stay. Thus, it is reasonable to assume an "ideal" situation where patient needs must be addressed within the week in which they arise. We use an offered load approach (i.e., we assume there are no constraints on care team capacity) and use the weekly realisations to estimate workload histograms and to determine reasonable capacity values. The capacity values that we derive assume that additional demand in a week must be addressed using overtime rather than fulfilling it in the future. In future work, we plan to consider a more realistic model where the care team prioritises certain patients and encounters based on the level of urgency related to the hospital readmission risk.

(3) Dedicated Staff: We assume that all staff types in the care team are solely dedicated to serving patients currently enrolled in the intervention. This is different from other hospital-based settings where high-risk, high need patients are only a subset of the total

patients seen by the staff members, such as nurses. Dedicated staffing models for patients with complex needs are common in interventions being piloted in the United States.

Section 6: Results

6.1. Independent workload estimation of each staff type

Figure 7 shows workload histograms for each staff type (data collected in steady state weeks across all the replications) based on 5000 replications of the simulation, along with means and 80th percentiles for a constant/deterministic number of patients ($\lambda = 3$ and $\lambda = 6$) being enrolled each week. Figure 8 shows the workload histograms for the case when the number of patients enrolled follows a Poisson distribution with a given mean ($\lambda = 3$ and $\lambda = 6$).

The workload histograms can be used for staffing purposes. For example, the care team may choose to staff at the 80th percentile. In the $\lambda = 3$ case and assuming a Poisson enrolment rate each week, this implies that the care team will plan for 46.58-h Community Health Worker's (CHW) time each week, which implies more than one CHW assuming a full-time workload of 40 h per week per CHW (or overtime); for the equivalent $\lambda = 6$ case the care team will plan for 87.5 h of CHW's time, which again implies more than 2 CHWs in the care team (or again, overtime). Notice that the 80th percentile values are slightly higher in the Poisson cases compared to the constant enrolment rate. The variability

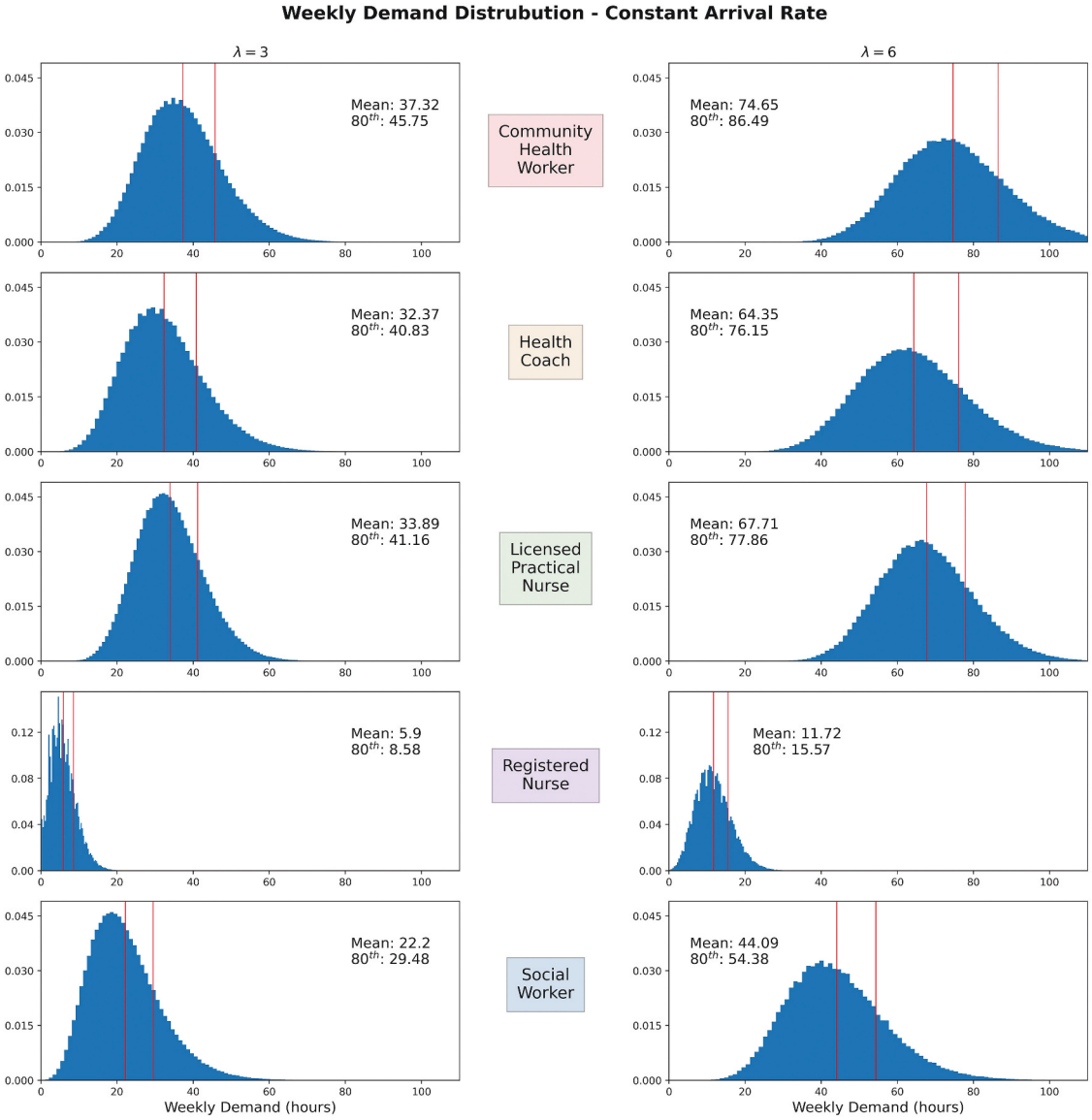

Figure 7. Workload histograms or weekly demand distributions (x axis in hours) for each staff type under constant weekly enrolment rates of 3 (left) and 6 (right).

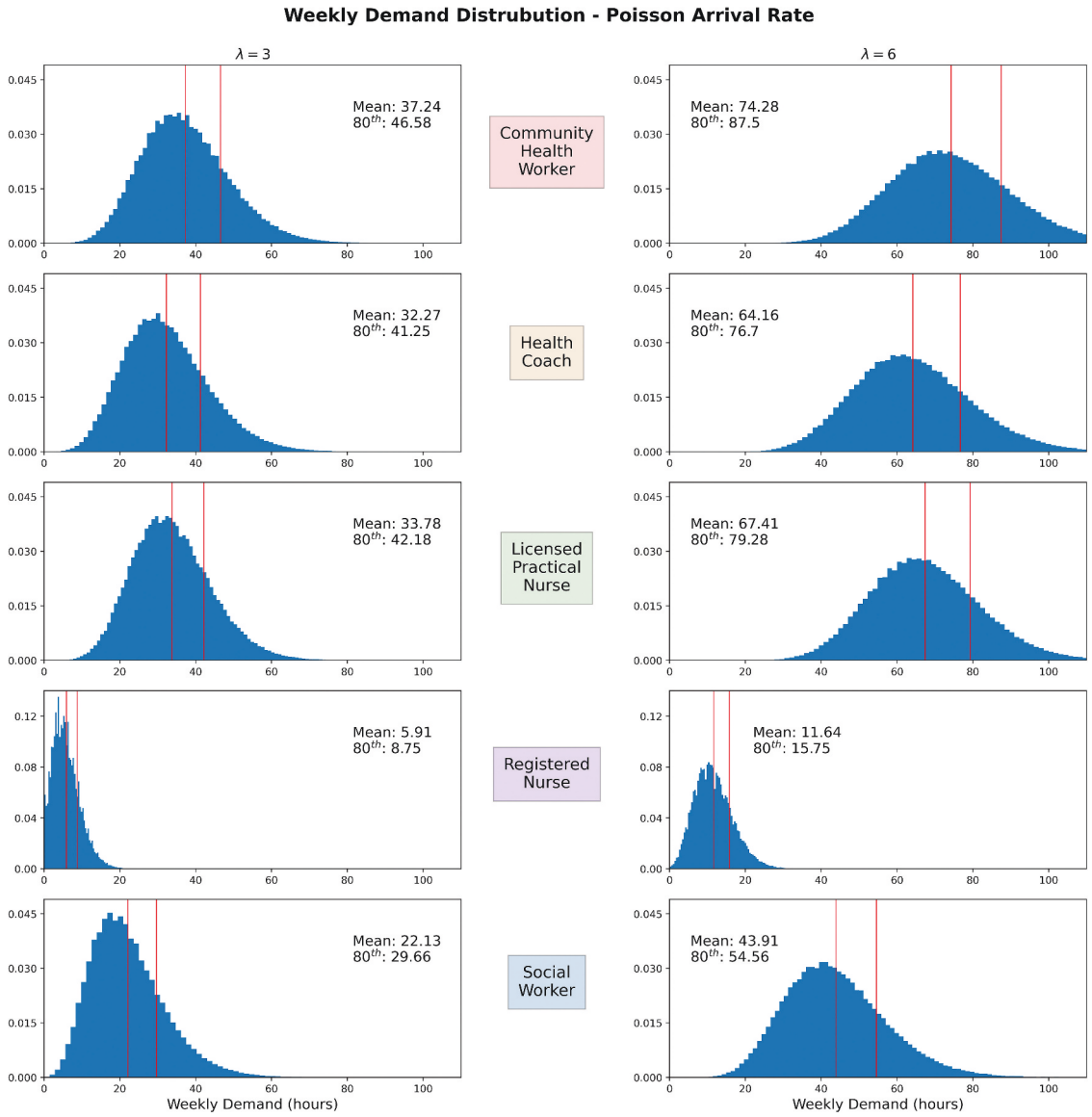

Figure 8. Workload histograms or weekly demand distributions (x axis in hours) for each staff type under poisson weekly enrolment rates of with means of 3 (left) and 6 (right).

in enrolments week to week creates some additional variability, however, from a practical point of view, the constant and Poisson cases seem to generate similar workload histograms. This suggests that the variability due to the time needed for encounters has a more significant impact than the variability due to weekly enrolments.

Workload histograms follow symmetric normal distributions closely and seem to scale up when the enrolment rate is doubled. The only exception is the RN histogram in $\lambda = 3$ case, which is truncated on the left by 0 – this is because RN hours per week are much lower than that for other staff types.

Care teams have to strike a balance between under-utilisation, when demand is below available staff hours, and utilisationover-utilisation, when demand is higher than available staff hours. The latter case implies inability to provide timely coordination and potentially increases the risk of patient hospitalisations and emergency events. We can construct a simple newsvendor model for this purpose. Newsvendor model is a single-period mathematical model used to determine optimal inventory levels in settings with fixed prices and uncertain demand for a perishable product. The modern formulation has been defined in Arrow et al. (1951). Newsvendor heuristics have been used in healthcare literature before. A recent example is Barz and Rajaram (2015) who calculated the critical fractile for different patient groups and prioritised them according to their net contributions.

While more sophisticated applications of the newsvendor model are possible, we use it here as a simple quantitative tool. We provide an illustrative example. Assuming the utilisationover-utilisation can be addressed by overtime, and an overtime cost (p) that is double the regular hourly rate (which corresponds to the cost of under-utilisation, c) for any staff member, we can plug

in the critical fractile formula for the newsvendor model to calculate the staffing level q:

$$q = F^{-1}\left(\frac{p-c}{p}\right) = F^{-1}\left(\frac{2-1}{2}\right) = F^{-1}(0.5)$$

Under the above assumptions, the critical fractile corresponds to the median. For CHW, the median is at 36 h per week and for LPN the median is at 32 h per week; these would therefore be the staffing levels for CHW and LPN respectively. For other values of p and c the staffing level q can be similarly calculated. An alternative is to calculate the expected weekly unsatisfied patient demand in hours and the unutilised hours as a function of the number of weekly hours for a staff type, for different values of q. This is shown in Figure 9. A care team could determine the value q based on the costs and implications that are best suited to their practice.

We see in Figure 9 that the distance between unutilised hours and unsatisfied demand is minimised at the median. One interesting observation is that because partial demand satisfaction is possible even for the demand levels above the capacity, the expected unsatisfied demand can be quite low. For example, if the LPN capacity is 32 h but the demand is 36 h, LPN can still satisfy 32 h of the demand. Only 4/36 or $1/9^{th}$ (about 11%) of the total demand was not satisfied. This means that even if we set the LPN capacity at the median of 32 h 88% of the weekly demand from the patient panel can be satisfied, and the unfulfilled demand could address through overtime or by moving non-urgent encounters to the next week. Such quantification will be important for care coordination teams that might be resource constrained.

6.2. Dependent workload estimation of home visits for LPN and CHW

Although it is easier to treat the demand for each staff type independently and calculate their workload separately, in reality, demand for each staff type is not independent. As explained above, staff members at Camden Coalition regularly work in teams of two – one nurse and one social worker/community health worker. This is particularly true when it comes to home visits. Regular home visits are a central feature of the Camden Coalition model. The Camden Coalition targets to have the first home visit of a patient within 5 days of discharge from hospital, and repeated home visits to the patients throughout their intervention. Home visits are the encounter type corresponding to the largest effort from staff,

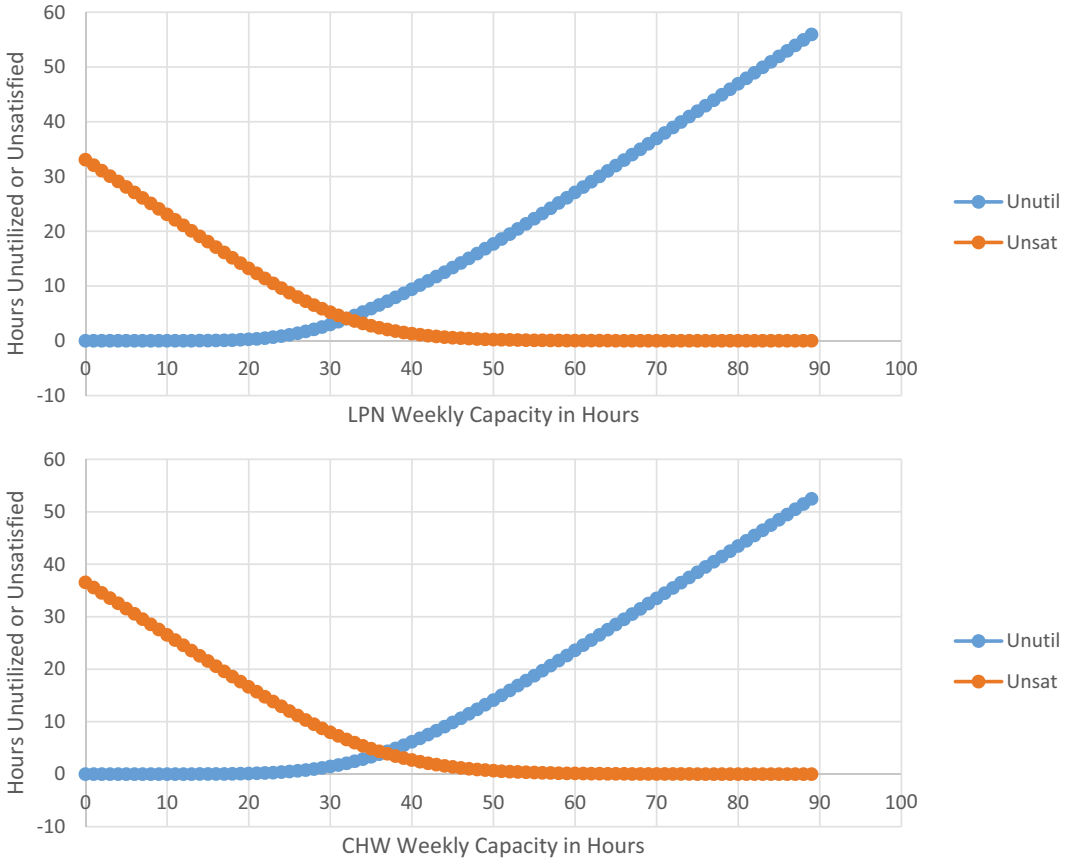

Figure 9. LPN and CHW expected unutilised unutilised hours (Unutil) and unsatisfied demand (Unsat) at different capacity levels. (Color Online).

accounting for 30.6% of total time effort (Martinez et al., 2019). Other care coordination programmes are likely to implement a similar care model.

For situations in which two or more staff types are involved in addressing the needs of a patient, calculating the staff weekly workload as a joint bivariate or multivariate distribution is more accurate. Fortunately, our simulation methodology and the level of granularity in our data allows for such calculations. Recall that $D_{s,t,k}$ represents the workload for staff type s in week t of replication k. We use $D_{s,t,k}$ across all steady state weeks in all replications to create workload histograms independently for each staff type. Now suppose that we are interested in estimating the bivariate weekly workload distribution for LPN and CHW. We would now use realisations of the pair $[D_{LPN,t,k}, D_{CHW,t,k}]$ across all steady state weeks in all replications to create a bivariate weekly workload histogram. If we further restrict the encounters to be only home visits, then we have the bivariate weekly workload histogram for LPN and CHW for home visits.

The key insight here is that since LPNs and CHWs frequently conduct home visits together, an increase (or decrease) in home visit workload in week for LPN also would mean an increase (or decrease) in home visit workload for the CHW, and vice versa. Thus, the home visit workloads for these staff types are correlated. This is what we mean by *dependent workload estimation*. In contrast, we can create a bivariate workload distribution by assuming independence. In the independent case, the probability that a LPN works h_1 hours in any given week and the CHW works h_2 hours is simply the product of the individual probabilities estimated separately based in histograms for LPN and CHW shown in Figure 7. That is, the home visit workloads for LPN and CHW have no correlation: an increase (or decrease) in one does not necessarily mean an increase (or decrease) in the other.

Figure 10 illustrates the difference between estimating the bivariate distribution under the independent and dependent assumptions. In the left panel, we see the bivariate histogram of number of home visit hours for LPN and CHW when we assume independence, and the in right panel we see the bivariate histogram when we assume dependence. These plots are calculated for LPN and CHW home visits under constant demand (3 patients enrolled per week).

Although the distributions look similar, there are a few key differences. The demand is more tightly focused around the mean and has a higher-valued single peak under the dependent distribution (right panel in Figure 10) when compared to independent bivariate distribution (left panel in Figure 10). This tighter distribution is due to the correlation in home visit hours that we discussed earlier.

The largest difference between the two distributions occurs when the when the LPN home visit demand is 13 h per week and CHW home-visit demand is 14 h per week. The dependent distribution estimates that this combination of home visit hours occurs 33.55% of the time, while the independent distribution estimates it occurs 26.42% of the time. Thus the independent distribution underestimates the joint demand by about 7.13% which can lead to an inaccurate staffing decisions. However, especially at higher capacities for the staff members, we find that the independent distribution is a good approximation of the dependent distribution.

We can use the bivariate cumulative distribution for home visits similar to how we used the independent cumulative distributions for all encounter types in a newsvendor model. If we assume that LPN and CHW have the same overtime and underutilisation cost ratio as before (2:1), we can reserve the capacity for home visits at the median. Using the bivariate cumulative distribution, the capacity values closest to the median are at 18 h for LPN and 15 h for CHW – those are the hours that could be reserved for weekly home visits based on the newsvendor model, when an average of 3 patients are enrolled in the program per week.

We can also similarly calculate the expected unutilised hours and unsatisfied demand per week for different combinations of LPN and CHW capacity levels as shown in Tables 3 and 4.

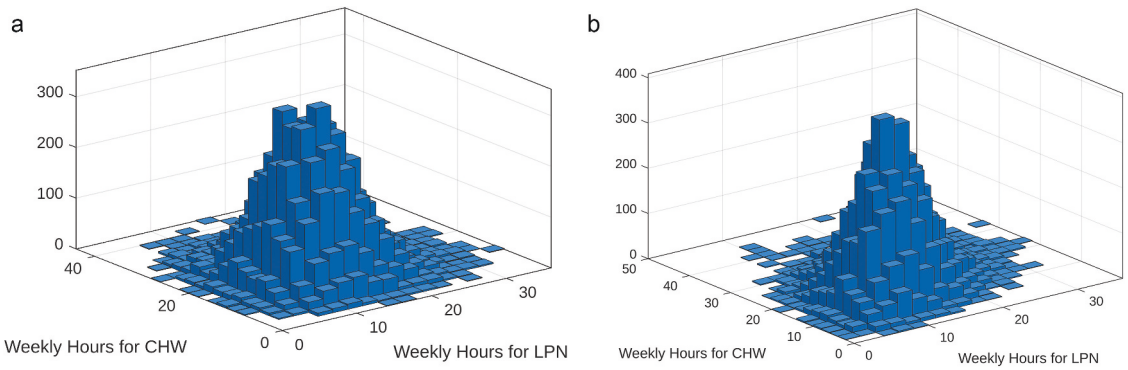

Figure 10. Independent (left) and dependent (right) bivariate histograms of home visits for LPN and CHW.

Table 2. LPN & CHW hours per week reserved for home visits vs. expected unutilised capacity in hours. Each cell contains two unutilised unutilised capacity values in hours: the first value for the LPN, the second for the CHW.

LPN\CHW	0	10	20	30
0	0,0	0,0.37	0,5.31	0,14.58
10	0.63,0	0.63,0.37	0.63,5.31	0.63,14.58
20	7.22,0	7.22,0.37	7.22,5.31	7.22,14.58
30	17.03,0	17.03,0.37	17.03,5.31	17.03,14.58

Table 3. LPN & CHW hours per week reserved for home visits vs. expected unsatisfied demand in hours. Each cell contains two unsatisfied home visit demand values in hours: the first value for the LPN, the second for the CHW.

LPN\CHW	0	10	20	30
0	12.97,15.45	12.97,5.82	12.97,0.77	12.97,0.04
10	4.43,15.45	4.43,5.82	4.43,0.77	4.43,0.04
20	0.66,15.45	0.66,5.82	0.66,0.77	0.66,0.04
30	0.02,15.45	0.02,5.82	0.02,0.77	0.02,0.04

To explain how to read and use above tables, we look at an example. If both LPN & CHW reserve 10 h per week for home visits, the weekly expected unutilised capacity will be 0.63 h for LPN and 0.37 h for CHW (Table 2) and the weekly expected unsatisfied demand will be 4.43 h for LPN and 5.82 h for CHW (Table 3). This is an unreasonable number of home visit hours unfulfilled and is likely to have an impact on patient outcomes. If CHW and LPN were staffed at 20 h each, the number of unsatisfied home visit hours goes down to 0.66 and 0.77-h respectively for the two staff types. On the flip side, the two staff types would have 7.22 to 5.31 h of unutilised home visit capacity. One thing to note is that while we are reserving the capacity specifically for home encounters (due to their importance in the care coordination intervention), unutilised capacity could still be used for other encounter types whereas unsatisfied demand will have negative connotations so it might make more sense to be more conservative. However, our newsvendor calculations above assume overtime costs for unsatisfied demand, so it might not be a significant concern if the staff types can be paid for additional hours beyond their allotted capacity.

6.3. 6.3. *Non-stationary enrolment rates*

In our analysis so far, we have assumed that while the number of enrolments from one week to the next can change (when sampled from the Poisson distribution), the mean weekly enrolment rates are unchanging or stationary. We now discuss how the impact of non-stationary enrolment rates can be analysed. Non-stationary enrolments are common in practice and arise for a variety of reasons. Care teams often start with a small number of patients, and as they learn the best practices of assisting patients and streamlining their processes, they enrol more patients and hire new staff members. Greater awareness of care coordination efforts in a community can also result in increased enrolments. A common problem in such situations is that as demand surges to a new equilibrium, the care team will need to adapt their staffing levels.

In this section, we model the case where a care team starts weekly enrolments that follows a Poisson distribution with a mean rate of 3 from week 1 to week 60. Starting in week 61, the mean weekly enrolment rate increases linearly over a 5-week period to a new mean of 5 in week 65. In other words, the distribution of weekly enrolments still follows the Poisson distribution, however the rate is non-stationary and represents a surge/increase over the current status quo. We assume that the mean weekly enrolment rate of 5 holds until week 120.

The simulation algorithm presented in Section 5 is modified to sample new enrolments each week based on the above non-stationary Poisson pattern. After experimenting with different number of replications to evaluate the stability of the results, we used 200 replications of the 120-week period to analyse the results. First, we plot how the average number of patients active in the program (i.e., estimate of E $[|W_t|]$ based on 200 replications)

Table 4. Percentiles of the RN weekly workload distribution for each of the four 20-week segments.

Percentile	Weeks 41–60	Weeks 61–80	Weeks 81–100	Weeks 101–120
10th	1.84	3.17	4.25	4.00
25th	3.42	5.33	6.42	6.25
50th	5.5	8.08	9.33	9.33
75th	8.16	11.42	12.83	12.66
90th	10.92	14.83	16.24	15.92

changes in the 120-week time period, see Figure 11 below. The average number of active patients also follows a non-stationary pattern. After the initial transient phase in weeks 1–35, the system reaches steady state for $\lambda = 3$. This steady state remains until week 61 when the mean weekly enrolment rate starts to increase. Although the mean weekly enrolments stop increasing in week 65, the number of active patients continues to rise in weeks 61–80 and stabilises partially in weeks 81–100 and more decisively in weeks 101–120. This demonstrates how an increase in weekly enrolment rate (from $\lambda = 3$ to $\lambda = 5$) in a 5-week period can create a much longer transient period of approximately 30–40 weeks.

To demonstrate how staffing levels should be adjusted, we divide the time horizon into 4 segments of 20 weeks each (weeks 41–60; 61–80; 81–100; 101–120). For each 20-week period, we estimate the workload histograms for each staff type based on the 200 replications. This gives us 4000 observations for each 20-week segment. As an example, Table 4 shows estimates of 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the registered nurse (RN) workload distribution in the each of the four 20-week segments. As expected, the percentiles in the transient 20-week segment lie in between the weeks 41–60 and 81–100 segments; and the percentiles in the last two 20-week segments are quite similar. These estimates can be used by care teams to gradually transition their capacities to meet the increase in demand.

By increasing the number of replications, the care team can also look at equivalent smaller time segments, if they need more precise short-term planning. For example, the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles for RN workload in weeks 61–65 (the weeks in which enrolments started rising) based on 200 replications are 2.33, 4.1, 6.58, 9.83 and 12.84 h, respectively. These numbers are within 2 h of the percentiles for weeks 41–60, and therefore suggest that an RN staffing needs to be increased only minimally in the surge weeks.

More complicated profiles than the linear demand surge we considered above can be similarly analysed. For instance, staffing estimates for a non-stationary pattern where certain months of the year have higher demand than others can also be quantified.

Section 7: Conclusions and future work

In summary, several insights can be obtained from our study. The first is that patient-level longitudinal data can be effectively used for aggregate capacity planning purposes, even when the number of patients in the dataset is small. Detailed event progression data for a small number of patients is a common feature of interventions involving patients with complex medical and social needs. This is because such patients only represent 1–5% of a population but can have a disproportionate impact on the health system, in terms of healthcare utilisation as well as costs.

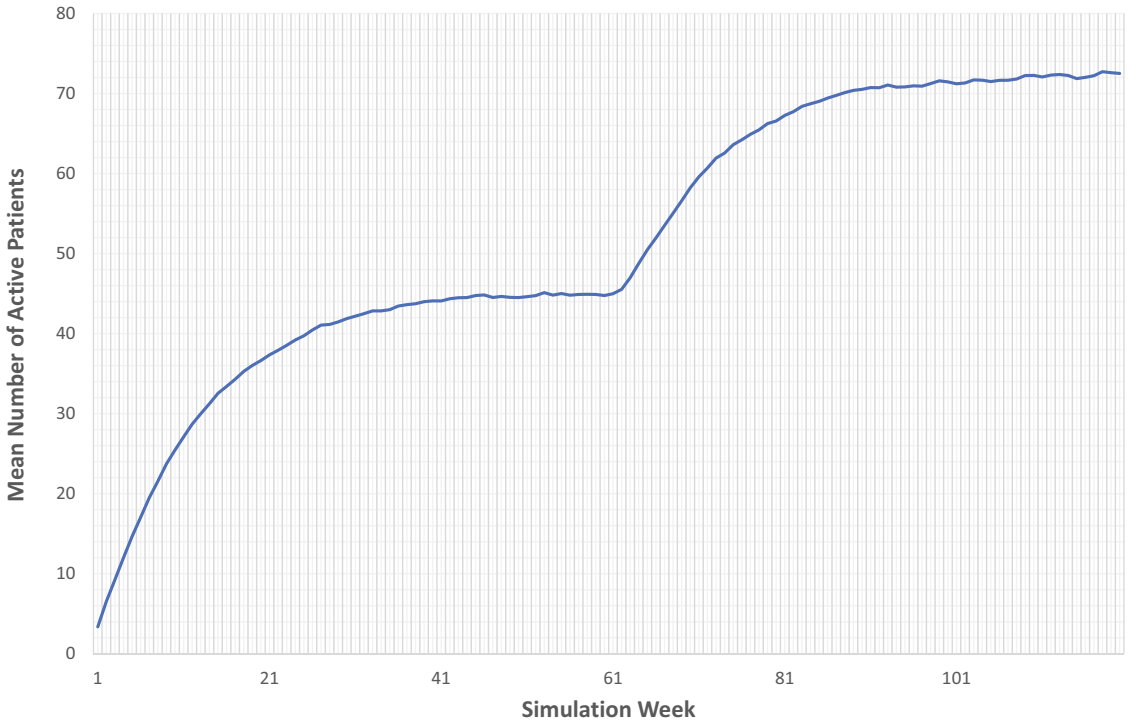

Figure 11. Average number of active patients in non-stationary case.

The second insight is that our sampling-based simulation algorithm can (a) implicitly model non-stationary/time-dependent features inherent in a patient's intervention and (b) explicitly model changes in the weekly enrolment rates. Traditional queueing network models are unable to capture such non-stationary dynamics. While discrete event simulation models of non-stationary queueing networks are possible, they require data for a large number of patients which are typically not available to care coordination programs.

Results based on our algorithm reveal several patterns. First, weekly workload distributions vary widely between staff types with registered nurses having the smallest effort while community health workers (CHWs) having the largest. As their name suggests, CHWs are recruited from the same community as the patients and are well versed with the available medical and social support systems. They therefore form a central pillar of the coordination effort, increasing the ability of the care team to effectively engage with patients. However, the pairing of different staff types for certain encounters is also a vital part of the care intervention. For example, a licenced practical nurse (LPN) and community health worker (CHW) often conduct home visits together. The LPN provides clinical expertise during the home visit while CHW helps connect the patient to agencies in the community that provide social support. Thus, both medical and social needs are coordinated together. Therefore, it is more accurate to consider the correlations in workload between staff types when making staffing decisions. Our results also demonstrate using the joint workload distribution involving multiple staff types provides a more accurate picture of staffing needs. Specifically, we found that for home visits, the LPN-CHW bivariate joint distribution is more tightly focused around the mean and has a higher-valued single peak compared to the LPN-CHW bivariate distribution created assuming independence.

Finally, our approach can also be used to create staffing plans when the mean weekly enrolment rate fluctuates. A key insight is that even short-term changes in mean weekly enrolments can result in long transient periods. Workload distributions for any small time segment in a transient period can be analysed with sufficient replications of our simulation algorithm, and can help care teams adjust their staffing levels. Such analysis is useful for planned increases or decreases in enrolment and seasonal patterns that recur each year. For unplanned changes, the nonstationary enrolments model could be used to generate what-if scenarios to scale up or scale down capacity levels for members of the care team.

We view the aggregate capacity planning approach in this study as a first step in the development of more detailed and sophisticated approaches. In future work, we plan create an analytical patient trajectory model using principled methods (such as maximum entropy estimation) that are able to overcome the limitations of sparse data and can capture the probabilistic trajectory of a patient beyond what is observed historically. Next, the inclusion of patient characteristics – combinations of specific medical conditions and social needs – that lead to greater effort or require higher priority would help with care team decision making. Finally, a model that explicitly considers patient wait times along with the prioritisation of patient types and patient encounters based on urgency (or risk of hospitalisation) would lead to policies on how best the care team should allocate their time.

Acknowledgments

This research was partly funded by U.S National Science Foundation grants NSF CMMI 1254519 and NSF CMMI 2212237. The views expressed in this paper are of the authors and not of the National Science Foundation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Agency for Healthcare Research and Quality (AHRQ). Multiple Chronic Conditions Chartbook. (2010). Medical expenditure panel survey data. Retrieved July 20, 2018, from https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/prevention-chronic-care/decision/mcc/mccchartbook.pdf

Ahmadi-Javid, A., Jalali, Z., & Klassen, K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258*(1), 3–34.

Arrow, K. J., Harris, T., & Marschak, J. (1951). Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, *19*(3) , 250–272.

Barz, C., & Rajaram, K. (2015). Elective patient admission and scheduling under multiple resource constraints. *Production & Operations Management*, *24*(12), 1907–1930.

Brandenburg, L., Gabow, P., Steele, G., Toussaint, J., & Tyson, B. J. (2015). Innovation and best practices in health care scheduling. *NAM Perspectives*. 5 2 https://doi.org/10.31478/201502g

Brown, R. S., Peikes, D., Peterson, G., Schore, J., & Razafindrakoto, C. M. (2012). Six features of medicare coordinated care demonstration programs that cut hospital admissions of high-risk patients. *Health Affairs*, *31*(6), 1156–1166.

Cady, R., Looman, W., Lindeke, L., LaPlante, B., Lundeen, B., Seeley, A., & Kautto, M. E. (2015). Pediatric care coordination: Lessons learned and future priorities. *OJIN: The Online Journal of Issues in Nursing*, *20*(3). https://doi.org/10.3912/OJIN.Vol20No03Man03

Campello, F., Ingolfsson, A., & Shumsky, R. A. (2017). Queueing models of case managers. *Management Science*, *63*(3), 882–900.

Cayirli, T., Veral, E., & Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, *9*(1), 47–58. https://doi.org/10.1007/s10729-006-6279-5

Chang, E., Ali, R., Seibert, J., & Berkman, N. D. (2023). Interventions to improve outcomes for high-need, high-cost patients: A systematic review and meta-analysis. *Journal of General Internal Medicine*, *38* (1) , 185–194.

Chow, V. S., Puterman, M. L., Salehirad, N., Huang, W., & Atkins, D. (2011). Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production & Operations Management*, *20*(3), 418–430.

Deglise-Hawkinson, J., Kaufman, D. L., Roessler, B., & Van Oyen, M. P. (2020). Access planning and resource coordination for clinical research operations. *IISE Transactions*, *52*(8), 832–849. https://doi.org/10.1080/24725854.2019.1675202

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, *40*(9), 800–819.

Helm, J. E., & Van Oyen, M. P. (2014). Design and optimization methods for elective hospital admissions. *Operations Research*, *62*(6), 1265–1282.

Hilton, R., Zheng, Y., Fitzpatrick, A., & Serban, N. (2018). Uncovering longitudinal health care behaviors for millions of medicaid enrollees: A multistate comparison of pediatric asthma utilization. *Medical Decision Making*, *38*(1), 107–119.

Howells, M., Andrew, L., & Gartner, D. (2022). Modelling the accessibility of adult psychology services using discrete event simulation. Proceedings of *Hawai'i International Conference on System Sciences (HICSS)*, Maui, Hawaii, USA, 3-7 January 2022

Keskinocak, P., & Savva, N. (2020). A review of the healthcare-management (modeling) literature published in manufacturing & service operations management. *Manufacturing & Service Operations Management*, *22* (1), 59–72.

Little, J. D. (1961). A proof for the queuing formula: L= λ W. *Operations Research*, *9*(3), 383–387.

Litt, J. S., & McCormick, M. C. (2015). Care coordination, the family-centered medical home, and functional disability among children with special health care needs. *Academic Pediatrics*, *15*(2), 185–190. https://doi.org/10.1016/j.acap.2014.08.006

Martinez, Z., Koker, E., Truchil, A., & Balasubramanian, H. (2019). Time and effort in care coordination for patients with complex health and social needs: Lessons from a community-based intervention. *Journal of Interprofessional Education & Practice*, *15*, 142–148.

Mattke, S., Mengistu, T., Klautzer, L., Sloss, E. M., & Brook, R. H. (2015). Improving care for chronic conditions. *RAND Research Report*, *5*(2).

Powers, B. W., Modarai, F., Palakodeti, S., Sharma, M., Mehta, N., Jain, S. H., & Garg, V. (2020). Impact of complex care management on spending and utilization for high-need, high-cost medicaid patients. *The American Journal of Managed Care*, *26*(2), e57–63. https://doi.org/10.37765/ajmc.2020.42402

Rossi, M. C., & Balasubramanian, H. (2018). Panel size, office visits, and care coordination events: A new workload estimation methodology based on patient longitudinal event histories. *MDM Policy & Practice*, *3*(2), 2381468318787188. https://doi.org/10.1177/2381468318787188