Soluzioni di *operations management* per il percorso ambulatoriale dell'ASST Nord Milano: tecniche e strumenti per la gestione delle code al CUP

Noemi Danzo, Bernardo Provvedi, Clara Carbone, Anna Prenestini*

L'attività ambulatoriale è uno dei fulcri della produzione delle aziende sanitarie e rappresenta uno dei principali momenti di contatto e cura tra i cittadini e i servizi sanitari. In questa ottica, è fondamentale per le aziende garantire un percorso e un'esperienza di cura che siano volti a ridurre le attese e a massimizzare il tempo a valore per l'utente. La ricerca si è posta l'obiettivo di individuare delle soluzioni organizzative e logistiche volte a ridurre le attese presso il *front office* centrale del Centro Unico di Prenotazione (CUP) del Presidio Ospedaliero Bassini dell'ASST Nord Milano, mettendo in pratica i principi e le analisi previste dalla teoria delle code. La metodologia utilizzata ha visto un utilizzo combinato di metodi qualitativi e quantitativi. Nello specifico è stato sviluppato un modello statistico che permette di individuare nelle diverse fasce orarie il numero di sportelli ottimale per minimizzare il tempo di attesa degli utenti. I risultati ottenuti sono stati quantificati con una riduzione del tempo medio di attesa presso il CUP centrale del 45%. Questo risultato è stato raggiunto attraverso diverse soluzioni operative, quali la riorganizzazione delle attività nelle diverse fasce orarie, la definizione e la diffusione di istruzioni operative per gli operatori di *front office*, nonché una serie di iniziative volte alla sensibilizzazione e alla formazione del personale coinvolto a diversi livelli.

Parole chiave: ospedale, azienda sanitaria, *front office*, CUP, logistica dei pazienti, percorso ambulatoriale, teoria delle code, tempi di attesa.

Operations management solutions for outpatients at ASST Nord Milano: Techniques and tools for queue management

The outpatient pathway is one of the most important pipelines for healthcare providers and represents the first point of contact between citizens and healthcare services. Therefore, healthcare organizations need to ensure a patient experience that aims to reduce waiting times and maximize the value-added time for the user. This research aims to identify orga-

SOMMARIO

* *Noemi Danzo*, Gestione Operativa, ASST Nord Milano.
Bernardo Provvedi, Gestione operativa, ASST Nord Milano.
Clara Carbone, ASST Santi Paolo e Carlo, Milano; già direttore SC Gestione Operativa ASST Nord Milano.
Anna Prenestini, Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano. Corresponding author.

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

nizational and logistics solutions to reduce queues at the central front office for admission, payment, and booking of the Bassini Hospital at ASST Nord Milano, applying the principles and analyses defined by the queuing theory. The methodology was a combination of qualitative and quantitative methods. Specifically, a statistical model was developed to identify the optimal number of front office operators for each hour to minimize users' waiting time. The results obtained can be quantified as a 45% reduction in the average waiting time at the Bassini Hospital front office. This result has been achieved through a number of operational solutions such as the reorganization of the activities along the different time slots, the definition and dissemination of procedures to front office staff, and a series of initiatives aimed at raising awareness and training operators at all level.

Keywords: Hospital, Healthcare organization, front office, Patient flow logistics, outpatient, queuing theory, waiting time.

Articolo sottomesso: 11/04/2024, accettato: 06/01/2025

1. La rilevanza dei percorsi ambulatoriali nelle aziende sanitarie

Le aziende sanitarie sono enti complessi e costantemente sfidati a garantire servizi sanitari di qualità in un contesto di scarsità di risorse. Per cercare un equilibrio tra queste istanze apparentemente contrapposte, le aziende sanitarie hanno sviluppato tecniche e professionalità volte a ottimizzare le risorse a disposizione, e quindi a efficientare i processi, mantenendo una grande attenzione al valore generato per gli utenti. Per questo motivo, è diventato cruciale sviluppare processi integrati che migliorino il flusso dei pazienti all'interno delle strutture, in modo da fornire trattamenti tempestivi e che massimizzino l'utilizzo delle risorse disponibili (Hall, Belson, Murali, & Dessouki, 2013).

In letteratura si fa riferimento alla logistica del paziente, nell'ambito delle teorie relative all'*operations management* nelle aziende sanitarie, quale branca di ricerca che si occupa della gestione dei percorsi fisici dei pazienti all'interno delle strutture sanitarie. In particolare, essa prende in considerazione l'intero percorso del paziente – dalle modalità di accesso alla struttura, all'assegnazione del paziente al percorso clinico-assistenziale più appropriato, fino alla fase finale di dimissione e gestione del post-acuto (Bensa *et al.*, 2008) – con l'obiettivo di massimizzare le parti a valore per l'utente e di ridurre al minimo i passaggi non a valore.

I percorsi ambulatoriali (o anche detti percorsi *outpatient*) sono dedicati agli utenti di servizi sanitari le cui condizioni di salute e i trattamenti diagnostici e terapeutici non necessitano di un ricovero ospedaliero di più giorni ma di accessi che si concludono al massimo entro l'arco della giornata (Carbone & Prenestini, 2015).

Con l'avanzare delle tecniche mediche e diagnostiche, oltre che delle possibilità tecnologiche, le attività ambulatoriali stanno diventando uno dei fulcri della produzione delle aziende sanitarie non solo per l'attività di diagnosi e indicazione terapeutica *one-shot* ma anche per l'erogazione di cicli di prestazioni sanitarie con accessi multipli del paziente, ciascuno dei quali si con-

clude nella giornata. Alla luce di questa situazione, le aziende sanitarie devono investire su metodi e strumenti che consentano, da un lato, di ottimizzare i percorsi nel senso di una maggiore fluidità e minimizzazione dei tempi di attraversamento, dall'altro, di migliorare l'esperienza dei sempre più numerosi utenti che saranno presenti – anche solo per qualche ora – all'interno della struttura.

Le teorie e le tecniche di *operations management* possono essere, quindi, la chiave per contribuire al miglioramento dei percorsi dei pazienti e per ottimizzare il funzionamento dell'area ambulatoriale. Durante la pandemia da Covid-19, per via delle esigenze di distanziamento e separazione dei percorsi, si è assistito a una sorta di ribaltamento dei principi guida della gestione operativa degli *outpatient* favorendo, in ragione della sicurezza, la dispersione spaziale e temporale degli utenti (per esempio, allungando i tempi tra una visita e quella successiva, garantendo così un numero inferiore di pazienti negli stessi spazi di attesa). Con la fine della situazione emergenziale, le strutture sanitarie sono tornate a porre maggiore attenzione su un maggiore efficientamento dei tempi e delle risorse da allocare alle attività del percorso ambulatoriale.

Dal punto di vista della letteratura scientifica manageriale sia italiana sia internazionale, i percorsi *outpatient* sono stati sinora meno studiati rispetto alle altre pipeline produttive aziendali (percorso chirurgico, percorso dell'emergenza urgenza ecc.). La maggior parte dei contributi scientifici internazionali è pubblicata su *journal* di carattere medico e si riferisce a specifici aspetti di miglioramento delle attività per la chirurgia ambulatoriale (Karimi *et al.*, 2020) o cicli di visite/prestazioni di una singola disciplina (Casey *et al.*, 2009). Alcuni degli studi più interessanti sull'*operations management* applicato al percorso *outpatient*, pubblicati su *journal* di carattere manageriale, riguardano l'applicazione di modelli di simulazione e/o l'adozione di strumenti di *lean management* per migliorare vari tipi di inefficienze, in particolare i tempi di attesa, la variabilità degli arrivi ecc. (Santibanez *et al.*, 2009; Gupta, 2023). Sempre attraverso sistemi di simulazione, Cayirli e colleghi (2006) mirano a ottimizzare la programmazione delle visite ambulatoriali. Recentemente, Fortineau e Roussel (2022) lamentano ancora una carenza di letteratura rispetto alla pianificazione dei flussi dei pazienti e con il loro studio propongono una metodologia, applicandola a tre casi di *outpatient clinic*. Nessuno degli studi empirici citati, tuttavia, fornisce una completa visione delle attività che può compiere un *front office* aziendale e che, se non ottimizzate, possono incidere negativamente su molti step del percorso del paziente ambulatoriale.

Nello specifico, il percorso del paziente ambulatoriale può essere scomposto in tre fasi fondamentali:

1) prenotazione;
2) erogazione del servizio (dall'accesso del paziente presso la struttura sanitaria, al momento del *check-in* da parte dei servizi di *front office*, sino all'effettiva erogazione della prestazione ambulatoriale da parte del professionista sanitario);
3) post-produzione (produzione del referto, qualora asincrono rispetto all'erogazione del servizio come

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

nel caso della diagnostica per immagini) e/o prenotazione di *follow-up* o successivi accertamenti.

Il presente lavoro pone l'attenzione sul percorso dei pazienti *outpatient* nelle fasi che riguardano gli adempimenti amministrativi svolti nell'ambito del Centro Unico di Prenotazione (CUP): si tratta, da un lato, dell'attività che comunemente viene indicata come il momento dell'"accettazione" del paziente e riguarda il riconoscimento del paziente e della sua prenotazione, con l'eventuale pagamento della prestazione (nella fase di erogazione del servizio); dall'altro, della possibilità dei pazienti di poter prenotare prestazioni e/o *follow-up* direttamente presso il CUP (nella fase di prenotazione pre e post-visita).
Per ottimizzare gli step gestiti dal CUP nel percorso *outpatient*, questo studio applica la "teoria delle code". Rispetto ai più sofisticati modelli di simulazione, la teoria delle code tradizionale utilizza modelli analitici statici e risulta particolarmente utile nel descrivere processi non troppo complessi e trovare soluzioni in tempi brevi, soprattutto in sistemi caratterizzati da risorse scarse. Sebbene la teoria delle code abbia avuto una più estesa applicazione in studi relativi a percorsi *inpatient* e in emergenza urgenza (Broyles & Cochran, 2007; Bruin *et al.*, 2007), almeno due interessanti studi hanno implementato tali modelli anche su percorsi *outpatient* ospedalieri in due differenti casi studi in altrettanti Paesi asiatici (Mital, 2010; Amzian & Hamdal, 2018; quest'ultimo solo a livello di *conference proceedings*). L'originalità di questo lavoro di ricerca è quella di aver testato e applicato un modello le cui potenzialità sono state ancora poco esplorate nel campo dei percorsi *outpatient* e, a differenza degli altri due studi sopra citati, in un contesto occidentale fondato su un Servizio Sanitario Nazionale.
Nei prossimi paragrafi, dopo aver presentato gli obiettivi del lavoro, verranno richiamati gli elementi principali della "teoria delle code" che saranno applicati al caso empirico: il CUP centrale del presidio ospedaliero (PO) Bassini dell'ASST Nord Milano. Per discutere il caso sarà affrontata dapprima l'analisi della situazione iniziale, in seguito la definizione della metodologia di costruzione del modello e, infine, i principali risultati ottenuti.

2. Oggetto e obiettivi del lavoro

Le fasi che coinvolgono il CUP nel percorso del paziente ambulatoriale rappresentano una sfida strategica per qualsiasi azienda sanitaria.
In primo luogo, i CUP rappresentano il biglietto da visita delle organizzazioni, la prima impressione che gli utenti avranno dell'azienda. Garantire una migliore esperienza agli utenti, limitando il tempo non a valore, costituisce uno degli elementi fondamentali per assicurare un percorso ambulatoriale di elevata qualità. Bleustein e colleghi (2014) hanno dimostrato l'esistenza di un forte legame tra i tempi di attesa e la soddisfazione dell'utenza, in particolare sulla percezione della qualità delle cure ricevute da parte dei medici.
In secondo luogo, ridurre i tempi di attesa per le fasi amministrative di prenotazione, accettazione e pagamento è essenziale per evitare possibili ritardi a cascata nell'erogazione delle visite mediche. La letteratura in materia è da sempre alla ricerca di modelli di *scheduling* volti a minimizzare il tempo

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

medico inutilizzato (Rockart & Hofman, 1969; Westeneng, 2007; Wijewickrama & Takakuwa, 2005). È quindi rilevante considerare che anche le fasi amministrative del percorso *outpatient* possono diventare un collo di bottiglia per le visite mediche in momenti di particolare affollamento del *front office*, creando attese e spreco di tempo anche per il personale medico, con conseguente capacità produttiva non utilizzata.
L'obiettivo del presente lavoro è quindi duplice:

1) costruire un modello analitico per la gestione delle code di utenti al CUP basato sulla "teoria delle code" e testarlo su un caso empirico (il CUP centrale del PO Bassini);
2) individuare le soluzioni organizzative e operativo-logistiche per ottimizzare le attività amministrative del percorso del paziente ambulatoriale.

La metodologia ha previsto:

- analisi qualitative, tramite visite sul campo e interviste agli operatori, e analisi quantitative, grazie ai dati estratti dall'applicativo elimina-code aziendale (si tratta del sistema che assegna un numero di chiamata, la relativa priorità e registra i diversi passaggi effettuati dall'utente) al fine di effettuare un'analisi precisa delle cause che portavano alla formazione di lunghe attese al CUP centrale del Bassini;
- la definizione e applicazione di un modello statistico basato sulla "teoria delle code" per definire i meccanismi di ottimizzazione dell'attività del *front office*.

3. Elementi di teoria delle code e fattori di influenza

I diversi contributi in letteratura sull'ottimizzazione della gestione degli appuntamenti e dei CUP si differenziano in due categorie per le tecniche di analisi adottate (Cayirli & Veral, 2009): da un lato, modelli di simulazione dinamici e, dall'altro, modelli statici di tipo descrittivo e analitico.
La teoria delle code è stata sviluppata da A.K. Erlang nel 1904 nel tentativo di determinare i requisiti di capacità del sistema telefonico danese, in modo da ridurne i problemi di congestione. Da allora, essa è stata applicata a una larga scala di aziende di servizi come banche, aeroporti, centralini telefonici, ospedali oltre che a sistemi di emergenza urgenza (Green, 2013).
In sanità, i modelli di teoria delle code (*queueing models*) possono aiutare a determinare livelli adeguati di personale, attrezzature e letti, nonché a prendere decisioni riguardo all'allocazione delle risorse e alla progettazione di nuovi servizi (Green, 2013). Per esempio, sono stati svolti numerosi lavori su:

- flussi in PS e quota di abbandono dei pazienti (Broyles & Cochran, 2007; Roche *et al.*, 2007);
- *scheduling* ed effetti degli interventi in emergenza su quelli elettivi (Agnihothri & Taylor,1991; Fiems *et al.*, 2007);
- fabbisogno di posti letto nei reparti (Bruin *et al.*, 2007);
- ridurre i tempi di attesa nelle fasi di accettazione (Bailey, 1953).

Questa branca di ricerca è utile per comprendere e descrivere il comportamento dei sistemi in cui si generano code (*queueing system*) in modo da

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

poterli governare in modo più efficiente; tre sono gli elementi che in genere caratterizzano i sistemi che generano code nel settore dei servizi (Hall, 1990):

- i fruitori (*customer*), ovvero le persone in attesa di ricevere un servizio, caratterizzati da un tempo di arrivo;
- gli operatori (*server*), ovvero le persone che forniscono il servizio, caratterizzati da un tempo di processo (o di servizio);
- la coda (*queue*), ovvero il gruppo di fruitori che attende di essere servito, caratterizzata da un insieme di regole che stabiliscono l'ordine con cui i *customer* vengono serviti (c.d. una *queue discipline*).

Il livello di coda che si genera è influenzato da diversi fattori, tra cui:

- la distribuzione degli arrivi degli utenti (*arrival patterns*), che rappresenta la frequenza di arrivo degli utenti, ossia l'intervallo di tempo che trascorre tra l'arrivo di un utente e il successivo;
- i tempi di servizio (*service times*), ovvero i tempi operativi necessari per svolgere il servizio;
- il fattore di utilizzazione (*utilization*), che esprime quanto efficientemente sono utilizzate le risorse, essendo definito come il numero di operatori occupati in un certo istante diviso il numero di operatori totali in quell'istante;
- la dimensione del sistema (*system size*), definita come il rapporto tra la domanda media e il tempo di servizio medio, essa è anche un determinante del numero di operatori necessari.

Nella teoria delle code il fattore di utilizzazione è una misura particolarmente rilevante.
Come illustrato in Fig. 1, il tempo di attesa aumenta con l'aumentare del

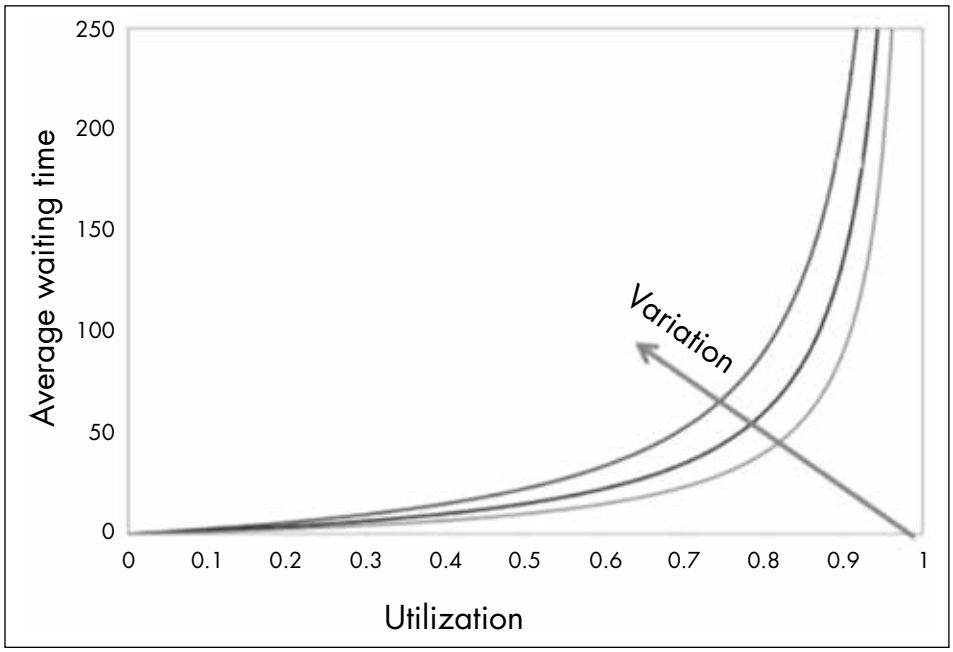


Fig. 1
Relazione fondamentale tra attesa e fattore di utilizzazione in un sistema a code (The Effects Of The Variation, n.d.)

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

fattore di utilizzazione in modo non lineare: l'attesa tende a infinito quando il fattore di utilizzazione si avvicina a uno (ossia tutti, o quasi tutti, gli operatori presenti sono occupati). Il punto esatto in cui la curva inizia a crescere esponenzialmente dipende da due fattori critici: la variabilità del sistema e la dimensione del sistema (Green, 2013).

Maggiore è la variabilità nel sistema, prima inizierà a flettere la curva e peggiore sarà il tempo di attesa per uno stesso livello di utilizzazione. Per quanto riguarda invece la dimensione del sistema si ha un effetto opposto: più è grande il sistema, più la curva inizierà a flettere vicino al 100%, quindi l'attesa risulterà minore per lo stesso livello di utilizzazione (Green, 2013).

In termini più analitici una formula molto utilizzata nella teoria delle code è quella del VUT che tenta di riassumere sinteticamente i principali fattori che influenzano i tempi d'attesa di un sistema. Essa permette infatti di evidenziare distintamente l'impatto dei tre fattori chiave: la variabilità (V), l'utilizzazione (U) e il tempo medio di processo (T). Di seguito è riportata la formula completa con relativa notazione[1]:

$$W_q = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u\sqrt{2(m+1)} - 1}{m(1-u)}\right)t_e$$

- W_q = tempo di attesa previsto in coda [*expected waiting time in the queue*];
- c_a = coefficiente di variazione dei tempi di arrivo;
- c_e = coefficiente di variazione dei tempi di processo;
- u = tasso di utilizzazione della postazione [*utilization of station*];
- m = numero di operatori paralleli in una postazione [*number of parallel-servers at a station*];
- t_e = tempo medio di processo [*mean processing time*].

Un'azienda sanitaria che voglia ridurre i tempi di attesa relativi a una sua attività può quindi agire su uno o più di questi fattori.

La variabilità esiste in tutti i sistemi produttivi e può avere un forte impatto sui livelli di performance. Molte teorie recenti (come la Lean o il Six-sigma) cercano proprio di spiegare questi impatti e di ridurne gli effetti negativi. Nella teoria delle code la variabilità viene calcolata sia per i tempi di arrivo degli utenti sia per i tempi di servizio e si misura utilizzando il coefficiente di variazione. Chand *et al.* (2008), con il loro studio, hanno dimostrato che identificando sistematicamente le fonti di variabilità in un processo e agendo su di esse è possibile raggiungere miglioramenti significativi nella performance del processo.

L'utilizzazione di un sistema o di un fattore produttivo è uno dei parametri storicamente più monitorati dalle discipline gestionali in sanità, si pensi a titolo di esempio alle saturazioni delle sedute operatorie o all'utilizzo efficiente delle grandi apparecchiature (TAC e Risonanze). A questo proposito è importante sottolineare come esista un *trade-off* tra tempi di attesa e saturazione, o in altri termini tra efficienza di flusso ed efficienza dei fattori produttivi (Modig & Åhlström, 2022). Come evidenziato dalla formula, ad alti tassi di utilizzazione corrispondono tempi di attesa crescenti. Si pensi a questo proposito come per servizi

[1] Le applicazioni pratiche della formula verranno discusse nel paragrafo seguente.

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

tempo-dipendenti si preveda sempre una capacità produttiva sovradimensionata (es. sale operatorie per l'urgenza). La scelta su quale combinazione di efficienza/fluidità del sistema si voglia attuare resta una decisione propria di ogni singola azienda. Il livello di utilizzazione critico, al di sopra del quale la coda degli utenti può andare fuori controllo, è diverso per ogni sistema, più il sistema è piccolo, più questo dovrà lavorare a un minore livello di utilizzazione per riuscire a servire tutti gli utenti con bassi tempi di attesa (Walley & Steyn, 2006).

Il tempo medio di servizio è il fattore che può apparire più intuitivo e facilmente interpretabile, minore è il tempo medio per gestire una richiesta, maggiore sarà il numero di utenti che potranno essere serviti in un dato lasso di tempo. In questo senso è importante notare come in servizi ad alta frequenza (alti volumi e breve durata) anche piccole variazioni sul tempo medio possono avere impatti significativi sull'accumulo di code. Inoltre, per ridurre il tempo medio di servizio e la sua variabilità si tende spesso a suddividere i flussi degli utenti sulla base di caratteristiche simili, in modo da avere risorse dedicate e specializzate. Questa suddivisione può però risultare controproducente: alcune risorse potrebbero ritrovarsi congestionate, altre libere senza lavoro da svolgere; alcuni pazienti potrebbero ritrovarsi nella coda sbagliata e attendere più del necessario dietro ad altri utenti con tempi molto più lunghi, o viceversa passare davanti a chi ha un'urgenza maggiore. Per queste ragioni, è spesso consigliabile mantenere un *pooling* di risorse non specializzate e tollerare un minimo di variabilità residua (Walley & Steyn, 2006).

Questi aspetti e le relative applicazioni pratiche verranno approfonditi nei paragrafi seguenti.

4. L'applicazione del modello della teoria delle code al PO Bassini dell'ASST Nord Milano

4.1. Il contesto e la situazione "*as-is*"

Il bacino di utenza dell'ASST Nord Milano è riconducibile al territorio a nord di Milano che fa riferimento a 6 Comuni (Cinisello Balsamo, Cologno, Cormano, Bresso, Cusano, Sesto San Giovanni) e conta quasi 270.000 abitanti. Esso è attualmente costituita da due presidi ospedalieri (Presidio Ospedaliero Edoardo Bassini di Cinisello Balsamo e Presidio Ospedaliero della Città di Sesto San Giovanni) e dalle sedi dei servizi socio-sanitari territoriali ubicati nel territorio a nord di Milano; oltre alla rete dei Poliambulatori territoriali dislocati in tutta l'area metropolitana di Milano. In particolare, il Presidio Ospedaliero Bassini è un ospedale generalista di medio-piccole dimensioni (275 PL), con un PS DEA di Primo livello e diverse specialità chirurgiche (urologia, oculistica, ortopedia ecc.). Annualmente l'azienda eroga circa 438.000 prestazioni ambulatoriali (escluso il laboratorio) di cui circa 195.000 vengono presso il PO Bassini.

Il CUP rappresenta il primo punto di contatto tra l'azienda e gli utenti che accedono alla struttura per effettuare prestazioni ambulatoriali. In questa sede vengono espletati i necessari adempimenti amministrativi, quali: l'accettazione, con l'eventuale pagamento della prestazione, e la prenotazione. Il servizio è articolato su un massimo di 10 sportelli al CUP Centrale, 4 sportelli al punto prelievi e 5 sportelli CUP dislocati in alcuni reparti. In media, al CUP Centrale vengono serviti circa 500 utenti al giorno.

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

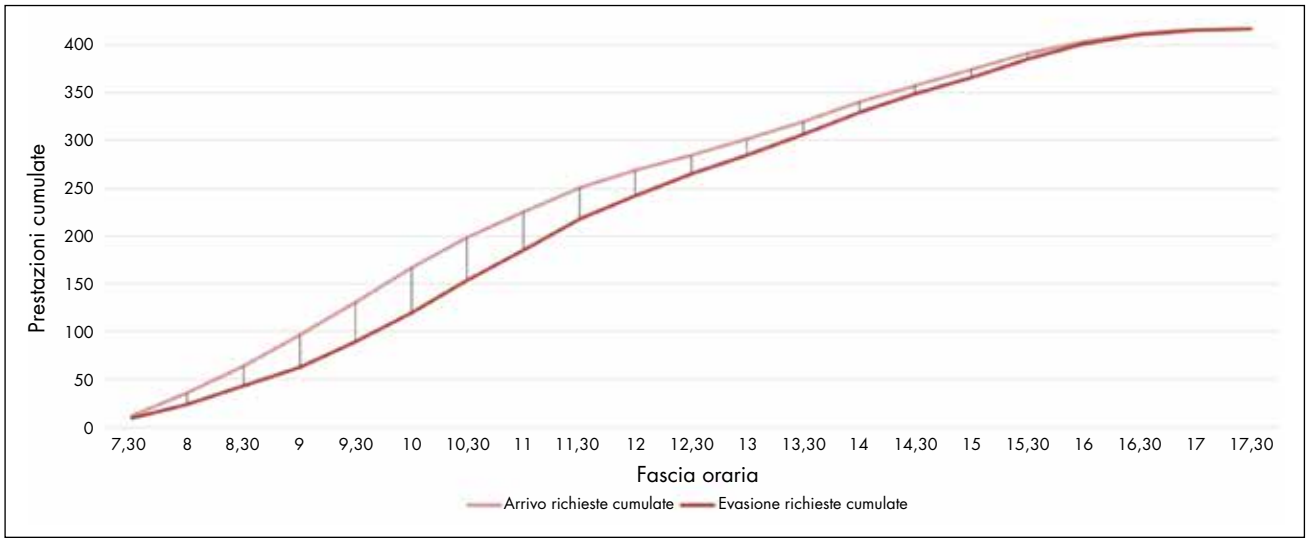
Durante la fase emergenziale dovuta al Covid-19 si era reso necessario decentrare il processo di accettazione degli utenti in più ambienti della struttura ("mini-CUP"), in modo da suddividere il flusso e garantire il distanziamento dei pazienti. Con il superamento della fase emergenziale, l'ASST Nord Milano è tornata a porre maggiore attenzione sull'uso efficiente delle risorse e quindi a svolgere la fase di accettazione presso il CUP Centrale dell'ospedale. In questa nuova fase di cambiamento sono emerse alcune criticità riguardo al metodo di gestione degli utenti al CUP Centrale, che portava alla formazione di lunghe code di pazienti in attesa e frequenti reclami presentati all'URP. Questa situazione generava quindi un impatto negativo sia sulla qualità del servizio percepita dagli utenti sia sull'utilizzo efficiente dei fattori produttivi.

Le principali cause delle lunghe attese, emerse attraverso le interviste con gli operatori e l'analisi dei dati estratti dal sistema di elimina-code, sono le seguenti:

- l'organizzazione dei turni giornalieri di *front office*: il numero di operatori disponibili agli sportelli nelle prime ore della mattina non era sufficiente a gestire il numero di utenti che si presentavano al CUP Centrale a richiedere un servizio; questa insufficienza portava a un inevitabile accumulo di coda e a lunghe attese, che si protraevano anche nelle ore successive, nonostante l'aumento di personale a disposizione;
- la mancanza di un metodo standard di chiamata degli utenti agli sportelli;
- la mancanza di un criterio univoco di utilizzo del *software* a disposizione da parte degli operatori, dovuta in parte alla mancanza di formazione;
- la difficoltà degli utenti nella selezione del servizio corretto da richiedere al totem del sistema elimina-code.

In Fig. 2 sono rappresentate la distribuzione cumulata degli arrivi al CUP Centrale (numero di richieste degli utenti) e la distribuzione cumulata degli utenti serviti dagli operatori (numero di evasioni di tali richieste).

Fig. 2
Distribuzione cumulata giornaliera di arrivo ed evasione delle richieste degli utenti per fascia oraria (dati estratti dal 18/05/2022 al 30/06/2022)



MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

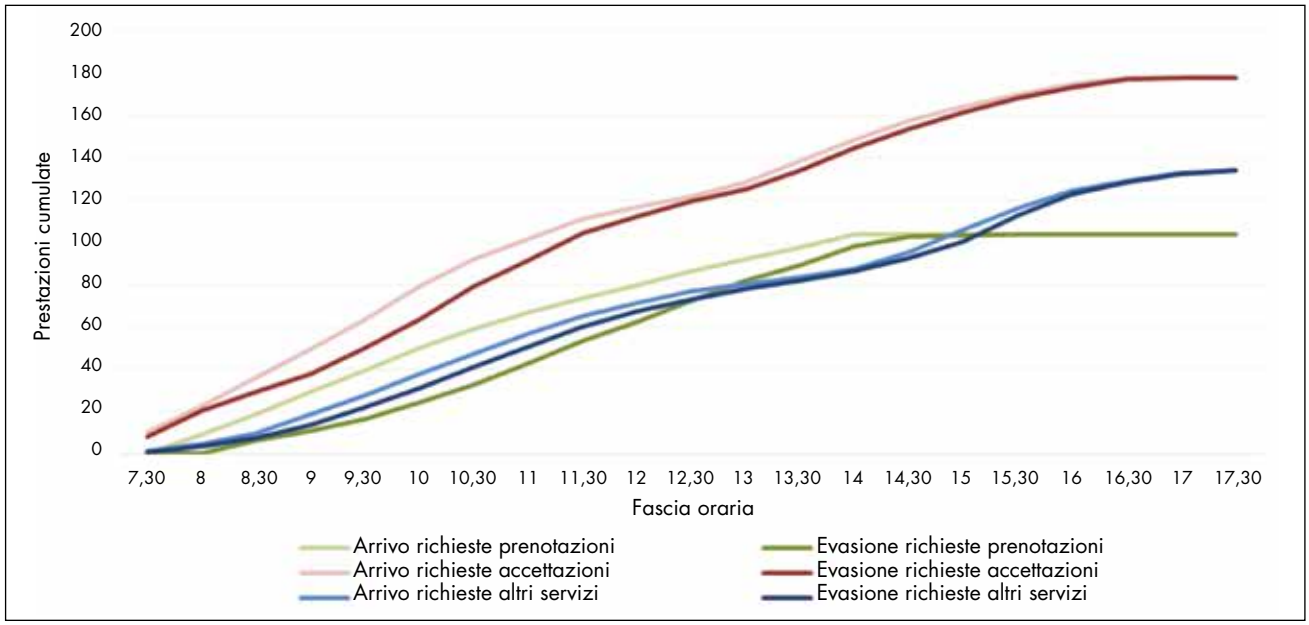
L'area che si trova tra le due curve è la rappresentazione grafica della coda di utenti che era mediamente presente in sala in quell'orario. Si può quindi osservare la presenza di un picco di attesa alle ore 10.00 con circa 50 utenti in coda (in una sala con una capienza massima di circa 40 persone, senza contare le restrizioni sull'utilizzo degli spazi dovuti al Covid-19). La coda tendeva a crearsi a inizio giornata, quando le richieste arrivavano più velocemente di quanto queste potessero essere evase e, non riuscendo a essere smaltita, tutti i nuovi utenti in ingresso finivano necessariamente ad accumularsi in coda; quest'ultima si esauriva solamente nel pomeriggio, quando anche le richieste degli utenti diminuivano.

In particolare, scomponendo per servizi il precedente grafico si nota come le maggiori discrepanze tra arrivi e richieste evase si concentravano sul servizio di prenotazione (si veda Fig. 3). Questo elemento risulta coerente con la necessità di dare precedenza alle accettazioni in modo da ridurre al massimo il rischio di ritardi nella visita ambulatoriale con conseguenti tempi morti per lo specialista.
Qui di seguito (Tab. 1), viene mostrato il tempo medio di attesa con i relativi indici di variabilità. Si osserva che il tempo medio di attesa era piuttosto elevato, di circa 30 minuti a utente, e che esso presentava un'elevata variabilità; entrambi i fattori sono strettamente collegati alla presenza continua di coda in sala.

Fig. 3
Distribuzione cumulata giornaliera di arrivo ed evasione delle richieste degli utenti per fascia oraria e tipologia di servizio (dati estratti dal 18/05/2022 al 30/06/2022)

Tab. 1 – Indici di variabilità del tempo di attesa (dati estratti dal 18/05/2022 al 30/06/2022)

Tempo d'attesa medio	00:28:27
Deviazione standard	00:30:22
Coefficiente di variazione (%)	107%



MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

Per comprendere meglio le dinamiche legate ai tempi d'attesa è opportuno scomporre l'analisi tenendo in considerazione le diverse tipologie di servizi offerti presso il CUP Centrale. In Tab. 2 si riporta il dettaglio dei servizi (con volumi giornalieri e tempi medi di attesa) presenti allo sportello CUP al momento dell'analisi.
I servizi che attirano il maggior numero di utenti sono il servizio di prenotazione ordinario per esterni e il servizio di accettazione, che quindi risultano essere i servizi più critici. Il servizio di prenotazione presentava un tempo di attesa molto elevato, di circa un'ora, e questo portava a continui reclami da parte dell'utenza e a un clima poco collaborativo in sala d'attesa. Il servizio di accettazione, invece, sforava spesso il tempo di attesa previsto di 15 minuti, valore identificato per limitare il rischio di ritardo del paziente alla visita medica (facendo così aspettare il medico, il cui tempo è una risorsa scarsa), con conseguente generazione di ritardo a cascata su tutti i pazienti successivi.

4.2. Costruzione del modello statistico

Per definire in modo oggettivo il fabbisogno di personale nelle diverse fasce orarie della giornata, è stato ideato un modello analitico, grazie all'utilizzo della teoria delle code, che permette di valutare i cambiamenti nel tempo medio di attesa previsto in coda al variare del numero di sportelli aperti. Questo modello, che è stato progettato internamente su Microsoft Excel, richiede dati di processo in input e permette di effettuare alcune simulazioni basiche di modifica del processo. La teoria delle code fa uso della legge di Little (*Little's law*), formulata nel 1961 da John D.C. Little. La legge di Little è una delle leggi fondamentali dei processi, essa descrive la relazione tra il *cycle time* (CT) di un processo (il

Tab. 2 – Dettaglio sul tempo di attesa medio per ogni tipologia di servizio (dati estratti dal 18/05/2022 al 30/06/2022)

Servizio	Utenti medi giornalieri	Tempo di attesa medio	Deviazione standard	Coefficiente di variazione
AA PRENOTAZIONI PRIMI ACCESSI (per prendere appuntamento i biglietti sono erogati dalle 8.30 alle 14.30)	106	01:01:44	00:35:28	0,57
PI ACCETTAZIONE PICCOLI INTERVENTI AMBULATORIALI, CITOLOGIA/ISTOLOGIA	8	00:13:33	00:14:13	1,05
CC ACCETTAZIONE IN DATA ODIERNA	176	00:19:16	00:17:50	0,93
DD ACCETTAZIONE PER APPUNTAMENTO IN DATA SUCCESSIVA	26	00:19:56	00:17:55	0,90
EE BAMBINO PRESENTE IN SALA (0-6 ANNI); DONNE GRAVIDE; INVALIDI; PS	28	00:05:08	00:05:02	0,98
FF VISITE PRIVATE (sportello dedicato dalle 15.30 alle 18.00)	23	00:08:53	00:09:14	1,04
LL PRENOTAZIONI POST-VISITA IN GIORNATA (dalle 8.30 alle 16.30)	43	00:24:32	00:22:17	0,91
BB PAZIENTI ESENTI; DH MEDICO; MEDICINA DEL LAVORO	20	00:09:37	00:10:49	1,12
Totale complessivo	431	00:28:27	00:30:22	1,07

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

tempo medio che un'unità impiega ad attraversare il processo), il WIP (*work in progress* o numero di unità in coda) e il Throughput (TH), ovvero il tasso medio di arrivo di nuovi elementi nella coda. La relazione è la seguente (Slack *et al.*, 2013):

$$\text{WIP} = \text{TH} \times \text{CT} \qquad (1)$$

Questa relazione è alla base delle formule utilizzate successivamente e rappresenta un caposaldo della teoria delle code. Con alcuni semplici passaggi è possibile calcolare il WIP previsto in coda considerando il tasso di arrivo e il tempo di attesa previsto in coda con la seguente formula:

$$\text{WIP}_q = r_a \times W_q \qquad (2)$$

Dove WIP_q rappresenta il numero di unità previste in coda [*expected work in progress in the queue*], W_q indica il tempo di attesa previsto in coda [*expected waiting time in the queue*] e r_a indica il tasso di arrivo degli utenti [*arrival rate* (*items per unit time*)].
I sistemi in cui si generano code possono avere caratteristiche diverse tra loro; in generale, essi sono caratterizzati da quattro parametri, utilizzati nella notazione di Kendall:

$$A/B/m/b$$

- *A*: la distribuzione dei tempi di arrivo (più precisamente *inter-arrival times*, ovvero il tempo trascorso tra gli arrivi);
- *B*: la distribuzione dei tempi di processo;
- *m*: il numero di operatori presenti in una postazione;
- *b*: il numero massimo di elementi ammessi nel sistema.

Le distribuzioni più comunemente utilizzate per descrivere i parametri *A* e *B* sono:

- la distribuzione esponenziale (o Markoviana), indicata con *M*;
- la distribuzione generale (per esempio la Normale), indicata con *G*.

Le due tipologie di sistemi più comuni e semplici sono i seguenti:

- *M*/*M*/*m*/-: distribuzione dei tempi di arrivo e di processo esponenziali con *m* operatori e nessun limite alla coda;
- *G*/*G*/*m*/-: distribuzione dei tempi di arrivo e di processo generali con *m* operatori e nessun limite alla coda.

Per ciascuna combinazione di parametri si possono utilizzare delle formule ricavate dalla legge di Little per calcolare il tempo medio di attesa previsto in coda. Per il caso *M*/*M*/*m*, più semplice, la formula è:

$$W_q = \frac{u\sqrt{2(m+1)} - 1}{m(1-u)} t_e \qquad (3)$$

Per il caso *G*/*G*/*m*, più comune, in quanto i tempi di processo raramente seguono una distribuzione esponenziale, occorre considerare anche una componente di variabilità in aggiunta alla formula (3) sopra presentata:

$$W_q = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u\sqrt{2(m+1)} - 1}{m(1-u)}\right) t_e \qquad (4)$$

Per applicare le formule della teoria delle code, è essenziale che gli arrivi e i tempi di servizio presentino una

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

natura stocastica, seguendo modelli probabilistici come le distribuzioni Poisson o esponenziali. È inoltre necessario garantire la stabilità del sistema, evitando che il tasso di arrivo superi la capacità di servizio disponibile. È fondamentale definire chiaramente il numero di operatori/sportelli e la struttura delle code, assicurandosi che il servizio non subisca interruzioni. Verificare attentamente queste condizioni nel contesto specifico è un prerequisito per ottenere stime affidabili dai modelli matematici.
In particolare, guardando le formule (3) e (4), una condizione fondamentale per l'utilizzo di questi modelli è che il tasso di utilizzazione (u) sia strettamente minore di 1. Nello specifico il tasso di utilizzazione è definito dalla seguente formula:

$$u = \frac{r_a}{r_e} = \frac{r_a\ t_e}{m}$$

dove: r_a rappresenta il tasso di arrivo degli utenti allo sportello; r_e è il tasso di processo con cui vengono gestite le pratiche al singolo sportello. Quest'ultimo può essere scomposto nel rapporto tra il numero di sportelli e il tempo medio di processo della singola pratica.
Da questa condizione ($u < 1$) deriva una prima importante considerazione: per poter applicare i modelli di teoria delle code è necessario avere un numero di sportelli aperti coerente con i tassi di arrivo e con i tempi medi di processo.
Nel caso in studio, la distribuzione dei tempi di arrivo e quella dei tempi di processo sono generali, quindi per i calcoli è stata utilizzata la formula (4). Dalla formula (4), presentata nei paragrafi precedenti come formula del VUT, si può dedurre che i tempi di attesa in coda aumentano con l'aumentare della variabilità, del fattore di utilizzazione e del tempo di processo.
Si deve sottolineare che i *queueing models* possono essere utilizzati sia in modo descrittivo sia in modo predittivo: in modo descrittivo, utilizzando i parametri dati in un certo istante, è possibile determinare per esempio quanto attenderà l'ultima persona che si è unita in coda; in modo predittivo, è possibile variare uno dei parametri e osservare come si modificherebbe il tempo di attesa in coda al variare di esso (Green, 2013).
Per questo lavoro il modello statistico è stato utilizzato inizialmente in modo descrittivo, per validarne l'uso confrontando il tempo di attesa risultante dal modello rispetto al tempo medio di attesa calcolato sui dati raccolti, e, in seguito, in modo predittivo. Utilizzando il modello in modo predittivo, è stato possibile introdurre diversi scenari di modifica del processo e osservare i risultati auspicabili, con tali modifiche, sui tempi di attesa.
In prima fase, è stato cruciale individuare delle fasce orarie chiave in modo da effettuare analisi puntuali e stabilire il numero adeguato di sportelli necessari in ciascuna di esse. Dopo diverse prove e valutazioni basate sull'analisi dell'andamento giornaliero degli arrivi degli utenti, si è optato per l'adozione di quattro specifiche fasce orarie:

- dalle ore 07.30 alle ore 09.00;
- dalle ore 09.00 alle ore 13.30;
- dalle ore 13.30 alle ore 15.30;
- dalle ore 15.30 alle ore 18.00.

I parametri in ingresso richiesti per l'utilizzo del modello sono:

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

- i tempi di processo, ovvero i tempi con cui vengono serviti gli utenti allo sportello, differenziati per tipologia di servizio;
- i tempi di arrivo degli utenti;
- i tempi di attesa;
- il numero di sportelli aperti.

Per la costruzione di questi parametri sono stati usati i dati estratti dal programma di elimina-code aziendale che permette di ricostruire il percorso del paziente dal momento di arrivo (selezione del servizio al totem) al momento di evasione della richiesta allo sportello. In particolare: il tempo di processo è stato calcolato come la differenza tra l'orario di evasione e quello di chiamata allo sportello; il tempo di arrivo è stato calcolato come il tempo che intercorre tra l'arrivo di un paziente e il successivo; il tempo di attesa rappresenta invece l'intervallo che intercorre tra l'arrivo del paziente e la sua chiamata allo sportello.
Rispetto ai tempi di processo, il tempo medio di ogni singola attività è stato aggiustato considerando un tempo di *set-up*[2] per tenere conto della disponibilità degli operatori. Il tempo di *set-up*, nel caso degli operatori del CUP, può essere interpretato come i tempi di pausa obbligatori previsti dal contratto. Questi tempi devono essere considerati per non rischiare di sovrastimare la capacità produttiva degli operatori.
Per confermare le assunzioni sui tempi di *set-up* e la validità complessiva del modello, sono state effettuate delle prove confrontando il tempo di attesa previsto in coda (stimato dal modello) con quello effettivamente registrato a sistema. Il confronto è riportato in Tab. 3.
A questo proposito possiamo osservare che il tempo risultante dal modello è molto simile a quello reale; lo stesso confronto è stato fatto utilizzando i dati del periodo temporale precedente, ottenendo sempre risultati simili. Inoltre, il fattore di utilizzazione risulta elevato, come potevamo aspettarci da quanto visto in precedenza nel terzo paragrafo, dato che a tassi di utilizzazione elevati corrispondono tempi d'attesa altrettanto elevati.
È importante sottolineare che, per effettuare questo confronto, non è stata considerata la suddivisione in fasce orarie. Infatti, trattandosi di un modello statico, la singola fascia oraria non terrebbe conto del WIP accumulato in precedenza e quindi i tempi risulterebbero minori rispetto a quelli reali. La suddivisione è invece utilizzabile, con eventuali distorsioni minime, per effettuare la previsione del numero di sportelli necessario in ogni fascia oraria, in quanto la volontà nell'utilizzo del modello è proprio quella di ridurre il WIP già dalla prima fascia oraria della giornata.
Con queste premesse, è stato possibile costruire il modello da utilizzare in modo predittivo. Trattandosi come abbiamo detto di un modello statico, si è cercato di renderlo semi-automatizza-

[2] Le formule e la notazione utilizzate per inglobare il tempo di *set-up* nel tempo di processo sono le seguenti (Wallace & Spearman, 2001):

$$t_e = t_0 + \frac{t_s}{N_s} \quad (5)$$

$$\sigma_e^2 = \sigma_0^2 + \frac{\sigma_s^2}{N_s} + \frac{N_s - 1}{N_s^2} t_s^2 \quad (6)$$

$$c_e^2 = \frac{\sigma_e^2}{t_e^2} \quad (7)$$

N_s = numero medio di unità processate
σ_s = deviazione standard del tempi di *set-up*
c_e = coefficiente di variazione dei tempi di processo con *set-up*
t_s = tempo medio del *set-up*
σ_0 = deviazione standard dei tempi di processo senza *set-up*
σ_e = deviazione standard dei tempi di processo con *set-up*
t_0 = tempo medio di processo senza *set-up*
t_e = tempo medio di processo con *set-up*

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

Tab. 3 – Confronto tra il tempo di attesa reale e quello previsto dal modello (dati dal 18/05/2022 al 30/06/2022)

r_a tasso medio di arrivo	t_0 tempo di processo medio iniziale	t_e tempo di processo medio aggiustato con *set-up*	*m* numero medio di sportelli aperti	*u* fattore di utilizzazione	W_q tempo di attesa previsto in coda (minuti)	*W* Tempo di attesa reale (minuti)
0,7	4,4	5,1	4	0,912	**29,07**	**28,45**

to e aggiornabile in maniera standard in caso di cambiamenti nel processo o di rilevante modifica delle ipotesi e dei parametri alla base di esso (come cambiamenti nei tassi di arrivo, nei tempi di processo, nei servizi effettuati ecc.). I passaggi effettuati sono stati trascritti per creare delle linee guida di costruzione del *database* e di implementazione del modello, mentre i dati nei fogli Excel sono stati collegati in modo che questi si aggiornino in automatico dopo aver effettuato l'inserimento nel nuovo *database* e aver compiuto alcuni semplici passaggi e valutazioni.

Si è partiti con la costruzione del *database*: è stata effettuata la pulizia dei dati, sono stati calcolati i parametri necessari come input al modello ed è stata creata la suddivisione nelle quattro fasce orarie giornaliere presentate all'inizio di questo paragrafo.

Costruito il *database*, il primo passaggio logico è stato quello di calcolare i parametri necessari raggruppando i dati per le diverse fasce orarie giornaliere. A fine esemplificativo, si riportano in Tab. 4 i parametri relativi ai tempi di arrivo raggruppati per fascia oraria.

Allo stesso modo, sono stati calcolati i parametri relativi ai tempi di processo, riportati in Tab. 5.

Tab. 4 – Parametri relativi ai tempi di arrivo utilizzati nel modello di teoria delle code (dati estratti dal 18/05/2022 al 30/06/2022)

Fascia oraria	t_a (minuti)	r_a ($1/t_a$)	c_a
07.30 – 09.00	1,3	0,8	1,6
09.00 – 13.30	1,1	0,9	1,9
13.30 – 15.30	1,6	0,6	1,1
15.30 – 18.00	2,6	0,4	1,5
tot	1,4	0,7	1,7

Tab. 5 – Parametri relativi ai tempi di processo utilizzati nel modello di teoria delle code (dati estratti dal 18/05/2022 al 30/06/2022)

Fascia oraria	t_0 (minuti)	$t_e = t_0 + \frac{t_s}{N_s}$ (minuti)	r_e ($1/t_e$)	c_e
07.30 – 09.00	4,4	4,4	0,2	0,9
09.00 – 13.30	4,5	5,2	0,2	1,4
13.30 – 15.30	4,4	5,2	0,2	1,4
15.30 – 18.00	3,8	3,8	0,3	1,2
tot	4,4	5,1	0,2	1,4

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

Sono stati calcolati anche i tempi di attesa e il numero medio di sportelli aperti nelle diverse fasce orarie, con il solo scopo di effettuare alcune verifiche e confronti.

A questo punto, per ogni fascia oraria, sono state svolte diverse ipotesi sul numero di sportelli aperti e si è osservato come sarebbero cambiati il tempo di attesa previsto in coda e il WIP previsto in coda al variare di essi. In Tab. 5 viene rappresentato il risultato ottenuto per la prima fascia oraria giornaliera. Guardando il fattore di utilizzazione, possiamo notare che questo è maggiore di 1 nei casi con 2 e 3 sportelli aperti; le formule della teoria delle code non possono essere applicate in questi casi. Il fattore di utilizzazione, per definizione, ha un campo di variazione che va da 0 a 1; se questo tasso è maggiore o uguale a 1, vuol dire che gli utenti arrivano più velocemente di quanto questi possano essere serviti e la coda aumenterà senza limiti. Inoltre, dato che nel modello il fattore $m\ (1 - u)$ si trova a denominatore, esso dovrà essere strettamente minore di 1, altrimenti risulterebbero tempi di attesa negativi o non calcolabili (nel caso in cui $u = 1$ il denominatore sarebbe 0, quindi la formula non potrebbe essere applicata). Considerando quindi solo i casi con un fattore di utilizzazione adeguato, si identifica il numero di sportelli aperti ipotetico che comporta tempi di attesa previsti in coda e WIP previsto in coda accettabili. In questo caso, avere almeno quattro sportelli aperti permetterebbe di ottenere un fattore di utilizzazione minore di 1 e tempi di attesa e WIP adeguati.

Il ragionamento appena effettuato va ripetuto per ogni singola fascia oraria. Prendiamo ora in considerazione l'ultima fascia oraria, il cui risultato viene mostrato in Tab. 7. Notiamo subito la differenza con la prima fascia oraria: in questo caso sarebbe sufficiente avere

Tab. 6 – Modello di teoria delle code per la scelta del numero di sportelli aperti adeguato alla fascia oraria 07.30-09.00 (dati estratti dal 18/05/2022 al 30/06/2022)

Fascia oraria	Numero di sportelli aperti ipotetico	$\frac{c_a^2 + c_e^2}{2}$	u	$u\sqrt{2(m+1)} - 1$	$m\ (1 - u)$	W_q (minuti)	WIP_q
07.30 – 09.00	2	1,7	1,7				
	3	1,7	1,1				
	4	1,7	0,8	0,7	0,7	7,5	5,7
	5	1,7	0,7	0,4	1,7	1,6	1,2
	6	1,7	0,6	0,2	2,7	0,6	0,4

Tab. 7 – Modello di teoria delle code per la scelta del numero di sportelli aperti adeguato alla fascia oraria 15.30-18.00 (dati estratti dal 18/05/2022 al 30/06/2022)

Fascia oraria	Numero di sportelli aperti ipotetico	$\frac{c_a^2 + c_e^2}{2}$	u	$u\sqrt{2(m+1)} - 1$	$m\ (1 - u)$	W_q (minuti)	WIP_q
15.30 – 18.00	1	1,9	1,5				
	2	1,9	0,7	0,6	0,5	8,6	3,3
	3	1,9	0,5	0,3	1,5	1,3	0,5
	4	1,9	0,4	0,1	2,5	0,3	0,1
	5	1,9	0,3	0,0	3,5	0,1	0,0

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

due sportelli aperti per ottenere un fattore di utilizzazione accettabile e un tempo di attesa previsto in coda e un WIP previsto in coda adeguati. Questa differenza è dovuta essenzialmente alla differenza nel tasso di arrivo e nel tempo di processo tra le due fasce orarie: in quest'ultima fascia entrambi i parametri sono nettamente inferiori rispetto a quelli della prima fascia, questo vuol dire che gli utenti arrivano più lentamente e vengono serviti più velocemente (probabilmente le richieste presentate in questa fascia sono più veloci da evadere in quanto il servizio di prenotazione ordinario, ovvero quello più lungo, termina alle 14.30).

Una volta effettuate queste valutazioni, per facilitare la successiva definizione dei turni orari, i risultati (ovvero i numeri di sportelli aperti prescelti) vengono sintetizzati in un unico schema, qui rappresentato in Tab. 8. Tutti i calcoli presentati finora si aggiornano in automatico nel momento in cui viene inserita una nuova base dati all'interno del file Excel; per questo singolo passaggio, sarà necessario inserire nella sezione apposita il numero di sportelli aperti scelto nella fase precedente e il resto dei dati verrà riportato in automatico dal *software*. Nello schema rappresentato di seguito, è stato inserito anche un confronto con il numero medio di sportelli aperti e con i tempi medi di attesa risultanti dalle analisi effettuate sui dati relativi al periodo considerato.

In questo caso, notiamo che, con il solo passaggio da tre a quattro sportelli aperti nella prima fascia oraria (nel caso in cui ci fossero risorse a disposizione), si otterrebbe un rilevante cambiamento nei tempi di attesa: il miglioramento beneficerebbe sia la prima fascia oraria sia le successive.

Infatti, la seconda fascia oraria, pur mantenendo invariato il numero di sportelli, otterrebbe anch'essa una riduzione nei tempi di attesa. I tempi di attesa reali risultano così elevati perché tengono conto del WIP accumulato in precedenza, il quale va a sommarsi ai nuovi utenti, che si presentano con un tasso di arrivo elevato. Il modello non considera l'influenza del WIP accumulato in precedenza sul tasso di arrivo, ma, data la netta riduzione che si otterrebbe nella prima fascia oraria, questo diventerebbe realmente ininfluente.

Il problema del WIP accumulato potrebbe ripercuotersi sulla terza fascia oraria, ma comunque in maniera molto più leggera rispetto a quanto già accade nella situazione reale. In aggiunta, nella terza fascia oraria il tasso di arrivo è nettamente inferiore, quindi il WIP precedente viene smaltito in tempi relativamente brevi.

Per quanto riguarda l'ultima fascia oraria, possiamo notare che i tempi risultanti dal modello sono simili a quelli reali. Infatti, in questa fascia, il tasso di arrivo è molto basso e il WIP accumulato in precedenza è poco influente.

Tornando al ragionamento sulla seconda fascia oraria, si potrebbe pensare di scegliere di avere sei sportelli aperti invece che cinque: in questo caso si avrebbero un tempo di attesa previsto in coda di 4,2 minuti in media e un WIP previsto in coda di 3,7 utenti in media. Ovviamente, la situazione va valutata attentamente considerando il *trade-off* tra efficienza di flusso ed efficienza delle risorse, oltre che le risorse effettivamente disponibili. Gli aggiustamenti sul numero di sportelli aperti possono

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

essere fatti sia aggiungendo personale, sia migliorando l'organizzazione dei turni nelle diverse sedi aziendali e gli spostamenti all'interno dei CUP di una stessa sede. Infatti, quando si ha necessità di personale al CUP Centrale, solitamente questo viene reperito effettuando degli spostamenti dagli altri CUP periferici della struttura, in cui il carico di lavoro diminuisce proprio nella seconda fascia oraria.

Tab. 8 – Sintesi del numero di sportelli aperti prescelto per fascia oraria e confronto con quelli medi del periodo (dati estratti dal 18/05/2022 al 30/06/2022)

Fascia oraria	Numero di sportelli aperti ipotetico scelto	W_q (minuti)	WIP_q	Numero medio di sportelli aperti nel periodo	*W* media nel periodo (minuti)	WIP_q teorico nel periodo
07.30 – 09.00	4	7,5	5,7	3	22,7	17,2
09.00 – 13.30	5	25,1	22,0	5	37,6	33,0
13.30 – 15.30	4	6,9	4,3	4	15,5	9,7
15.30 – 18.00	2	8,6	3,3	2	8,4	3,2

4.3. Simulazioni di modifica del processo

In fase di analisi sono stati testati diversi scenari al fine di trovare la soluzione ottimale che bilanci riduzione dei tempi d'attesa e utilizzo efficiente delle risorse.

Per simulare alcune modifiche di processo, sono stati presi in considerazione tre parametri governabili dall'organizzazione, ovvero il tempo di processo, la variabilità di processo e il tasso di arrivo. In particolare, sono state valutate tre situazioni:

- scenario 1: l'impatto di una riduzione minima dei tempi medi di processo, del 2%;
- scenario 2: l'impatto di una riduzione minima della variabilità di processo, del 5%.
- scenario 3: una modifica nel tasso di arrivo degli utenti, spostando nel *database* l'orario di inizio del servizio di prenotazione ordinario dalle 08.30 alle 09.30.

L'impostazione del modello permette di testare a livello teorico i diversi scenari, modificando semplicemente i parametri in ingresso, al fine di individuare le soluzioni che ottimizzano maggiormente il percorso *outpatient* prima di una loro eventuale implementazione concreta. Di seguito vengono mostrati i risultati di sintesi per ogni alternativa.

Scenario 1: riduzione del tempo di processo. In Tab. 9 viene mostrato il risultato ottenibile sui tempi di attesa con un'eventuale riduzione del 2% dei tempi di processo confrontandolo con il risultato ottenuto dal modello senza questa riduzione. Ciò che risulta sorprendente è quanto una variazione così piccola (5,3 secondi) nel tempo medio di processo possa avere un impatto significativo sui tempi di attesa complessivi. Per illustrare questo punto, prendiamo in considerazione la seconda fascia oraria: mantenendo costante il numero di sportelli aperti e ridu-

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

cendo il tempo medio di processo di circa 5 secondi, si otterrebbe una riduzione di circa 5 minuti nei tempi medi di attesa. Questo dimostra come anche modifiche relativamente piccole nei tempi di processo possono avere un impatto notevole sulla soddisfazione dell'utente e sull'efficienza complessiva del sistema. Per intervenire sui tempi di processo le aziende possono utilizzare alcune leve come, per esempio, l'ottimizzazione e la standardizzazione delle procedure, la formazione degli operatori, la riorganizzazione dei task all'interno del processo.

Tab. 9 – Sintesi del numero di sportelli aperti prescelto per fascia oraria per il caso generale, confronto dei tempi di attesa con una riduzione del 2% del tempo di processo rispetto al caso base (dati estratti dal 18/05/2022 al 30/06/2022)

	t_e minore del 2% (riduzione di 5,3 secondi)			Caso generale base (senza riduzione)		
Fascia oraria	Numero di sportelli aperti ipotetico scelto	W_q (minuti)	WIP_q	Numero di sportelli aperti ipotetico scelto	W_q (minuti)	WIP_q
07.30 – 09.00	4	6,4	4,8	4	7,5	5,7
09.00 – 13.30	5	19,8	17,4	5	25,1	22,0
13.30 – 15.30	4	6,2	3,9	4	6,9	4,3
15.30 – 18.00	2	7,8	3,0	2	8,6	3,3

Scenario 2: riduzione della variabilità di processo. In Tab. 10 viene mostrata l'ipotesi di una possibile riduzione del 5% nella variabilità di processo, quantificabile con una riduzione di 0,07 nel coefficiente di variazione. Anche in questo caso notiamo una possibile riduzione nei tempi di attesa, seppur minore rispetto a quella ottenibile con la riduzione del tempo medio di processo. In particolare, questo specifico scenario si concentra sulla possibilità di ridurre la variabilità del tempo di evasione delle richieste agli sportelli CUP, al netto della variabilità naturale derivante dalle singole operazioni da effettuare. Lo scenario 1 e 2 nella realtà sono necessariamente collegati: ridurre la variabilità nel processo standardizzando e velocizzando alcune procedure permetterebbe di ridurre anche i tempi operativi e quindi di combinare i benefici ottenibili sul tempo medio di attesa.

Tab. 10 – Sintesi del numero di sportelli aperti prescelto per fascia oraria per il caso generale, confronto dei tempi di attesa con una riduzione del 5% della variabilità di processo rispetto al caso base (dati estratti dal 18/05/2022 al 30/06/2022)

	c_e minore del 5% (riduzione di 0,07)			Caso generale base (senza riduzione)		
Fascia oraria	Numero di sportelli aperti ipotetico scelto	W_q (minuti)	WIP_q	Numero di sportelli aperti ipotetico scelto	W_q (minuti)	WIP_q
07.30 – 09.00	4	7,3	5,5	4	7,5	5,7
09.00 – 13.30	5	24,2	21,2	5	25,1	22,0
13.30 – 15.30	4	6,5	4,1	4	6,9	4,3
15.30 – 18.00	2	8,3	3,2	2	8,6	3,3

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

Scenario 3: modifica dell'orario di inizio del servizio di prenotazione (tasso di arrivo). Dall'analisi dei dati il servizio di prenotazione ordinario è risultato essere quello con maggiori criticità legate ai tempi di attesa. Infatti, tale servizio presenta un elevato tempo di processo che porta alla formazione di lunghe code. In Tab. 11 viene mostrato il risultato ottenibile sui tempi di attesa con lo spostamento dell'orario di inizio del servizio di prenotazione ordinario dalle ore 08.30 alle ore 09.30. Questo scenario prende in considerazione l'ipotesi di allineare temporalmente il momento di maggior afflusso di utenti (fasce orarie con maggior numero di richieste) con il momento di maggior capacità recettiva (fasce orarie con maggior numero di operatori disponibili allo sportello).

Tab. 11 – Sintesi del numero di sportelli aperti prescelto per fascia oraria per il caso generale, confronto dei tempi di attesa con lo spostamento dell'orario di inizio delle prenotazioni ordinarie rispetto al periodo di riferimento (dati estratti dal 18/05/2022 al 30/06/2022)

Spostamento orario di inizio delle prenotazioni ordinarie				Dati del periodo di riferimento		
Fascia oraria	Numero di sportelli aperti ipotetico scelto	W_q (minuti)	WIP_q	Numero medio di sportelli aperti nel periodo	*W* media nel periodo (minuti)	WIP_q teorico nel periodo
07.30 – 09.00	3	18,0	11,7	3	22,7	17,2
09.00 – 13.30	6	4,7	4,2	5	37,6	33,0
13.30 – 15.30	4	6,9	4,3	4	15,5	9,7
15.30 – 18.00	2	8,6	3,3	2	8,4	3,2

Quest'ultimo scenario è stato poi implementato scegliendo di spostare l'orario di inizio del servizio di prenotazione ordinario dalle ore 08.30 alle ore 10.00. Questa modifica ha permesso di ridurre i tempi di attesa giornalieri al CUP Centrale, agendo solamente sul tasso di arrivo degli utenti, quindi senza necessità di risorse aggiuntive. In questo modo, è stato possibile garantire tempi di attesa più brevi nonostante la limitata disponibilità di operatori, gestendo in maniera più flessibile le risorse attraverso lo spostamento degli operatori dagli altri CUP aziendali (es. laboratorio) nei momenti di maggiore necessità. Questa soluzione mira a ottimizzare l'allocazione delle risorse e a ridurre i tempi di attesa complessivi, offrendo un servizio più efficiente e soddisfacente per gli utenti. La verifica dell'efficacia di questa modifica è stata monitorata attraverso un successivo rilevamento dei dati e un confronto con i risultati ottenuti prima dell'implementazione del nuovo orario di servizio (Tab. 12).

Tab. 12 – Tempo di attesa medio prima e dopo lo spostamento dell'orario di inizio del servizio di prenotazione

Tempo di attesa medio un mese prima della modifica	Tempo di attesa medio un mese dopo la modifica	Differenza
00:30:45	00:16:40	– 00:14:05

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

4.4. Implementazione di soluzioni organizzative e logistico-operative

Il presente studio è stato realizzato tenendo in considerazione le specificità strutturali e organizzative dell'ASST Nord Milano. Infatti, il contesto aziendale è tipico di aziende sanitarie di medio-piccole dimensioni, con spazi ridotti (es. sale di attesa) e un grado di maturità digitale contenuto (es. assenza di totem informatizzati per la gestione autonoma di accettazioni/prenotazioni – c.d. *self check-in*). Questi aspetti hanno reso particolarmente rilevante il tema delle attese all'interno dei presidi aziendali e hanno sollecitato l'individuazione di soluzioni alternative.

Dal punto di vista organizzativo e logistico-operativo sono state intraprese diverse azioni. In primo luogo, è importante sottolineare che, per mantenere gli operatori informati e allineati relativamente ai motivi che hanno portato all'avvio del progetto, nonché al suo stato di avanzamento, sono state predisposte fin dall'inizio delle riunioni periodiche di formazione e di aggiornamento. In queste riunioni, venivano comunicate le principali criticità riscontrate e le possibili vie percorribili per contrastare tali problematiche, ascoltando anche il punto di vista degli operatori. Questa pratica contribuisce a una comunicazione chiara e trasparente, favorendo una migliore comprensione e partecipazione del personale.

Come primo passo è stato riorganizzato il metodo di gestione dei turni giornalieri degli operatori di *front office,* cercando di allineare il numero di sportelli aperti nelle diverse fasce orarie a quello previsto dal modello e quindi al flusso delle richieste degli utenti. Grazie allo spostamento dell'orario di inizio del servizio di prenotazione è stato possibile mantenere invariato il numero di operatori presenti nelle prime ore della mattina e gestire i turni successivi con un aumento del numero di operatori adeguato all'aumentare del flusso degli utenti. Questa strategia ha contribuito a evitare sovraccarichi e sottoutilizzo del personale, massimizzando così l'efficienza operativa complessiva del *front office* e garantendo allo stesso tempo un servizio più rapido e di qualità per gli utenti.

Per quanto concerne la standardizzazione dei metodi di chiamata degli utenti agli sportelli e delle procedure di lavoro, è stato necessario inizialmente compiere una revisione e semplificazione dei tasti selezionabili al totem da parte dell'utenza. Alcune delle modifiche effettuate sono le seguenti:

- è stato creato un unico tasto per le accettazioni ordinarie, accorpandole con il servizio di accettazione per piccoli interventi, in quanto il servizio reso all'utenza è lo stesso e, in entrambi i casi, necessita di essere evaso tempestivamente;
- è stato creato un tasto a parte per la medicina del lavoro e per il DH medico, che in precedenza si trovavano accorpati con le accettazioni per i pazienti esenti;
- il tasto per i pazienti esenti è stato prima inserito a parte come "accettazione pazienti esenti" e poi rimosso, a seguito delle continue lamentele del personale, in quanto esso veniva erroneamente selezionato anche per effettuare prenotazioni;
- è stato previsto un tasto per chi ha già pagato il ticket tramite PagoPA e necessita solo la stampa della ricevuta, in modo da incentivare tale attività.

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

Eseguito questo passaggio di semplificazione dei tasti selezionabili da parte degli utenti al totem, si è passati a stabilire delle priorità di chiamata. Le priorità, concordate con il referente del CUP aziendale, sono state poi inserite nel sistema di elimina-code utilizzato dagli operatori e uniformate tra i diversi sportelli. In questo modo, effettuando una chiamata "standard" vengono chiamate per prime alcune categorie di utenti preimpostate (Categorie prioritarie, Visite del medico competente e Stampa ricevuta PagoPA) e in seguito il restante degli utenti sulla base del tempo di attesa. Fino a quel momento, gli operatori avevano sempre optato per eseguire una chiamata specifica, selezionando il servizio da un apposito menu a tendina. La selezione del servizio da chiamare avveniva sulla base delle indicazioni comunicate nel corso della giornata o dei reclami presentati da parte dell'utenza. In generale, un paio di operatori erano dedicati al servizio di accettazione, mentre gli altri si suddividevano in maniera casuale i restanti servizi, senza tener conto dei tempi di attesa.

Con questa presentazione, è stato suggerito di utilizzare principalmente la funzione di chiamata standard, per garantire tempi di attesa livellati tra i vari servizi, mantenendo però un paio di operatori dedicati al servizio di accettazione, in quanto per tale servizio vanno garantiti tempi di attesa limitati.

In aggiunta, per assicurare tempi di attesa inferiori ai 15 minuti per l'accettazione, è stato stimato che il numero di utenti in coda non debba mai essere superiore a 10, perciò, nel caso in cui il numero risulti superiore, è stato chiesto agli operatori di intervenire per farlo scendere nuovamente. La stima del numero massimo di utenti in coda è stata ottenuta grazie alla formula del WIP, ovvero la formula (2) vista in precedenza, utilizzando il tasso medio di arrivo giornaliero e i 15 minuti di attesa massimi come parametri. Sono stati scelti 15 minuti come limite massimo di attesa per l'accettazione in modo da assicurare lo svolgimento della prestazione sanitaria in orario, ipotizzando che un paziente si presenti in media tra i 20 e i 30 minuti in anticipo rispetto all'orario della prenotazione.

Infine, sono state create delle linee guida contenenti tutte le indicazioni presentate, in modo da distribuirle e renderle consultabili in ogni momento da parte degli operatori.

Le condizioni organizzative che hanno facilitato la realizzazione del presente studio all'interno dell'azienda possono essere sintetizzate in tre elementi:

1) la presenza di un nucleo storico di operatori di sportello, che ha reso più semplice la formazione e il coinvolgimento degli operatori più giovani;
2) una stretta collaborazione tra la Struttura Complessa di Gestione Operativa e la Struttura CUP Aziendale, dovuto a uno storico di lavori precedenti che hanno consolidato le relazioni e integrato le modalità di lavoro;
3) la collaborazione con istituti universitari che ha permesso il coinvolgimento di tirocinanti e ricercatori per l'approfondimento di tematiche innovative.

4.5. Considerazioni sul modello statistico

Come già enunciato nei paragrafi precedenti, il modello, costruito con l'uti-

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

lizzo della teoria delle code, è statico, ovvero opera fintanto che sono validi i parametri in ingresso. Data la staticità del modello, nel caso in cui alcuni parametri come il tasso di arrivo o di processo dovessero variare, questi andrebbero aggiornati per poter eseguire nuove valutazioni sulla necessità di effettuare delle modifiche di processo o di reperire risorse aggiuntive.
La progettazione del modello con un'architettura che facilita gli aggiornamenti è perciò cruciale per garantire la sua rilevanza e affidabilità nel tempo. Consideriamo per esempio un'ipotesi di offerta che incide sulla domanda, vale a dire che i pazienti, osservando un processo di prenotazione più scorrevole, potrebbero sentirsi maggiormente incentivati a prenotare presso l'azienda. Tale circostanza potrebbe comportare un incremento del tasso di arrivo e, di conseguenza, una necessità di aumentare il numero di operatori presenti agli sportelli.
La flessibilità del modello nel recepire e adattarsi a scenari come quello appena descritto è fondamentale per garantire che le previsioni siano allineate con la realtà operativa. La capacità di modificare agevolmente le variabili coinvolte, come il tasso di arrivo, riflette la praticità e l'efficacia del modello nel gestire cambiamenti nelle condizioni del processo. Questa flessibilità è fondamentale in contesti in cui i cambiamenti sono inevitabili, consentendo al modello di rimanere un assetto affidabile e di valore nel lungo periodo.
Dalla necessità di flessibilità e adattamento del modello nasce anche l'esigenza di ridurre al minimo l'intervento manuale necessario per gli aggiornamenti. Ciò implica la definizione di procedure standardizzate e il collegamento dei dati in modo intelligente. Le linee guida sulla costruzione del *database* forniscono una struttura chiara per l'inserimento e la gestione dei dati. Inoltre, il modello è stato implementato in modo da consentire l'aggiornamento automatico dei dati nei fogli Excel dopo l'inserimento nel nuovo *database* e l'esecuzione di alcune valutazioni di routine.
La documentazione dettagliata dei passaggi eseguiti non solo serve come riferimento per gli utenti coinvolti nel processo, ma costituisce anche una base solida per future ottimizzazioni o modifiche al modello.
Considerati i risultati conseguiti con lo sviluppo del modello statistico per il PO Bassini, l'intenzione è di estenderne l'applicazione ad altri contesti aziendali, quali il CUP del PO di Sesto San Giovanni e i CUP dei Poliambulatori territoriali. Per adattare il modello a tali contesti, è necessario effettuare una rielaborazione dei dati, poiché il sistema di elimina-code è diverso in queste strutture. Inoltre, è fondamentale condurre osservazioni sul campo per definire con precisione le variabili da considerare, inclusi elementi come il tempo di pausa degli operatori o altre variabili specifiche del processo in questione.

5. Conclusioni

Nel presente studio, è stata impiegata la teoria delle code per sviluppare un modello semi-automatico e aggiornabile. Utilizzando dati di input facilmente accessibili ed elaborabili, il modello identifica il numero ottimale di operatori necessari agli sportelli nelle diverse fasce orarie per garantire tempi di attesa accettabili e ridurre l'accumulo di coda in sala d'attesa. Questo modello consente anche simulazioni di modifiche al processo, valu-

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

tando l'efficacia di diverse azioni. Una specifica azione correttiva è stata implementata modificando l'orario di inizio del servizio di prenotazione, ottenendo una riduzione significativa dei tempi di attesa, pari al 45%, senza richiedere risorse aggiuntive. L'efficacia di questa azione è stata costantemente monitorata, evidenziando un impatto positivo sui tempi di attesa senza compromettere il numero di utenti serviti.

Da un punto di vista qualitativo, è stato cruciale uniformare le procedure di lavoro e implementare un metodo univoco di chiamata degli utenti agli sportelli. La partecipazione attiva degli operatori è stata fondamentale, coinvolgendoli fin dall'inizio e condividendo analisi, criticità e azioni da intraprendere. Questo approccio ha contribuito a mitigare resistenze e favorire un ambiente aperto al cambiamento e di collaborazione e di apprendimento reciproco tra unità operative di staff e di line.

Inoltre, da un punto di vista manageriale, questo progetto ha dato la possibilità di diffondere la cultura del dato come strumento per la modifica dei processi organizzativi. A partire dall'analisi dei dati è stato possibile sensibilizzare gli operatori alle nuove modalità di lavoro e ridurre le resistenze al cambiamento basando le decisioni su fenomeni oggettivi e non sulle semplici impressioni.

In conclusione, l'approccio adottato ha affrontato aspetti tecnici e organizzativi del processo, producendo risultati positivi sia in termini di efficienza operativa sia di clima lavorativo. La riduzione dei reclami da parte degli utenti e l'ottimizzazione dell'ambiente di lavoro evidenziano il successo complessivo delle azioni implementate e un aumento generale di fiducia nei confronti dell'azienda e del personale che vi opera. Il passo successivo, una volta consolidate le innovazioni organizzative introdotte, sarà quello di sensibilizzare e supportare l'utenza (anche quella più anziana) a interfacciarsi con percorsi più veloci attraverso sistemi di *self check-in*, soprattutto per quei cluster di pazienti esenti per i quali non è previsto un pagamento ma soltanto una semplice operazione di accettazione. L'azienda dovrà, quindi, dotarsi anche di sistemi di accoglienza che possano ben orientare gli utenti all'interno della struttura investendo sui diversi mezzi di comunicazione e segnaletica.

6. Ringraziamenti

Le autrici e l'autore intendono rendere omaggio alla memoria della dott.ssa Elisabetta Fabbrini, allora Direttore Generale dell'AUSL Nord Milano, per aver creduto e dato valore alla gestione operativa aziendale, sostenendo questo progetto e autorizzando l'utilizzo dei dati ai fini di ricerca.

Inoltre, ringraziano l'attuale direzione strategica aziendale, rappresentata dal Direttore Generale Tommaso Russo, Direttore Sanitario Valentina Bettamio, Direttore Socio Sanitario Barbara Mangiacavalli e Direttore Amministrativo Vincenzo Centola per aver continuato a riconoscere l'importanza dei progetti di *operations management* e di logistica dei pazienti, garantendo il supporto necessario per lo sviluppo di nuove attività di analisi, sperimentazione e ricerca.

BIBLIOGRAFIA

Aziati A.H.N., Hamdan N.S.B. (2018). Application of queuing theory model and simulation to patient flow at the outpatient department. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, March, 3016-3028.

Agnihothri S.R.,Taylor P.F. (1991). Staffing a centralized appointment scheduling department in Lourdes Hospital. *Interfaces*, 21: 1-11.

Bailey N.T.J. (1954). Queuing for medical care. *Applied Statistics*, 3: 137-145.

Bensa G., Prenestini A., Villa S. (2008). La logistica del paziente in ospedale: aspetti concettuali, strumenti di analisi e leve di cambiamento. In: CERGAS – Bocconi (a cura di). *Aziendalizzazione della Sanità in Italia: Rapporto OASI 2008* (pp. 327-364). Milano: Egea.

Bleustein C., Rothschild D.B., Valen A., Valaitis E., Schweitzer L., & Jones R. (2014). Wait Times, Patient Satisfaction Scores, and the Perception of Care. *The American Journal of Managed Care*, 20(5): 393-400. – https://europepmc.org/article/med/25181568.

Broyles J.R. and Cochran J.K. (2007). Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression. In *Proceedings of the 2007 Industial Engineering Research Conference*, 613-618.

Bruin A.M., Koole G.M., Visser M.C. (2005). Bottleneck analysis of emergency cardiac in-patient flow in a university setting: an application of queueing theory. *Clinical and Investigative Medicine*, 28: 316-317.

Cayirli T., Veral E., and Rosen H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9(1): 47-58.

Cayirli, T. and Veral, E. (2009). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12: 519-549. DOI: 10.1111/j.1937-5956.2003.tb00218.x.

Casey J.T., Brinton T.S., and Gonzalez C.M. (2009). Utilization of lean management principles in the ambulatory clinic setting. *Nature Reviews Urology*, 6 (3): 146.

Fiems D., Koole G. and Nain P. (2007). *Waiting times of scheduled patients in the presence of emergency requests*. – http://www.math.vu.nl/~koole/articles/report05a/art.pdf, accessed August 6, 2007.

Fortineau V., Russell L. (2022). Planning Patient Journeys in Outpatient Hospitals to Support the Ambulatory Shift. In: Chaabane S., Cousein E., Wieser P. (a cura di). *Healthcare Systems: challenges and opportunities*, pp. 195-208. Wiley.

Green L. (2013). Queueing Analysis in Healthcare. In: Hall R. (eds). *Patient Flow. International Series in Operations Research & Management Science*, 206: 361-384. Springer, Boston, MA. DOI: 10.1007/978-1-4614-9512-3_15.

Gupta P., Pranjal P., Sachan A., Prasad P. (2021). Managing arrival variability in healthcare services: case of an out-patient department. *International Journal of Productivity and Quality Management,* 34(2): 135-157. DOI: 10.1504/IJPQM.2021.118380.

Hall R. (1990). *Queueing Methods for Service and Manufacturing.* Upper Saddle River, NJ: Prentice Hall.

Hall R., Belson D., Murali P., & Dessouky M. (2013). Modeling Patient Flows Through the Health care System. In: Hall R. (eds). *Patient Flow* (pp. 3-42). *International Series in Operations Research & Management Science*, 206. Springer, Boston, MA. DOI: 10.1007/978-1-4614-9512-3_1.

Karimi A., Sepehri M.M., Yavari E. (2020). A simulation model approach to decrease the length of stay of patients undergoing cataract surgery. *Perioperative Care and Operating Room Management*, 21, 100133.

Mital K.M. (2010). Queuing analysis for outpatient and inpatient services: a case study. *Management Decision*, 48(3): 419-439. DOI: 10.1108/00251741011037783.

Modig N., Åhlström P. (2022). *This is lean. Come risolvere il paradosso dell'efficienza* (Ed. italiana a cura di Faorlin A.). Milano: Guerini Next.

Roche K.T., Cochran J.K. and Fulton I.A. (2007) Improving patient safety by maximizing fast-track benefits in the emergency department – a queuing network approach. In *Proceedings of the 2007 Industrial Engineering Research Conference*, pp. 619-624.

Rockart J. F., and Hofmann P. B. (1969). Physician and Patient Behavior under Different Scheduling Systems in a Hospital Outpatient Department. Medical Care, 7(6): 463-470. – http://www.jstor.org/stable/3763110.

Santibáñez P., Chow V.S., French J. *et al.* (2009). Reducing patient wait times and improving resource utilization at British Columbia Cancer

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677

Agency's ambulatory care unit through simulation. *Health Care Management Science*, 12: 392-407.

Slack N., Brandon-Jones A., Johnston R. (2013). *Operation Management*, 7th Edition (pp. 361-367). Pearson Education.

Wallace J.H., Spearman M.L. (2001). *Factoryphysics:foundations of manufacturing management,* 2nd Edition. Boston: Irwin/McGraw-Hill.

Walley P., Steyn R. (2006). Managing variation in demand: lessons from the UK National Health Service. *Journal of Healthcare Management*, 51(5): 307-319.

Westeneng J.B. (2007). *Outpatient appointment scheduling: an evaluation of alternative appointment systems to reduce waiting times and underutilization in an ENT outpatient clinic*. University of Twente, Enschede, The Netherlands School of Management and Governance.

Wijewickrama A., Takakuwa S. (2005). Simulation analysis of appointment scheduling in an outpatient department of internal medicine. *Proceedings of the Winter Simulation Conference. Orlando, FL, USA, 2005*, pp. 10. DOI: 10.1109/WSC.2005.1574515.

MECOSAN – ISSN 1121-6921, ISSNe 2384-8804, 2024, 132 DOI: 10.3280/mesa2024-132oa20677