IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

Patient-Centric Multi-Prescription Intake Architecture: Algorithmic Temporal Grouping and Low-Latency Data Federation in Specialty Pharmacy

Neeraj Gaddam

Independent Researcher, Leander, Texas, USA

ABSTRACT: Specialty prescriptions seldom arrive alone. A patient starting a combination regimen is usually referred with several prescriptions that reach the pharmacy minutes or hours apart through e-prescribing, fax and provider portals. Intake systems that open one case per prescription split each such episode into several cases, repeat demographic and benefit verification, let conflicting records coexist and consume technician time that is increasingly scarce. This paper presents a patient-centric intake architecture that consolidates the prescriptions of an episode into one case before work begins. It has three parts: a federated entity-resolution service that links documents to a patient across channels with a probabilistic scorer and an ambiguity margin; a keyed session-window grouping algorithm with a bounded hold, an urgency bypass and late attachment; and straight-through processing (STP) rules evaluated on the consolidated case at the time each decision is made. The design is evaluated in a discrete-event simulation of synthetic referral streams (30 replications of about 17,736 prescriptions each) against per-prescription, tumbling-window, session-window and immediate-release baselines. With a 20-minute gap and a 60-minute cap, it created 48.3% fewer cases and needed 35.7% fewer technician minutes than per-prescription intake, raised the straight-through share from 21.1% to 35.0%, and held prescriptions for a median of 20 minutes (maximum 60 minutes) while urgent prescriptions were not held. Federated resolution split 0.17% of multi-prescription episodes, against 20.1% for an exact name-and-birth-date key, and the ambiguity margin prevented wrong links between twins in a stress test. Workload and timing parameters are assumed rather than measured; across the sensitivity scenarios the design still needed 29.7% to 40.7% fewer technician minutes than per-prescription intake.

KEYWORDS: Specialty pharmacy, prescription intake, event correlation, session windows, entity resolution, data federation, straight-through processing, healthcare workflow automation

I. INTRODUCTION

Specialty medications treat complex, chronic and often rare conditions such as cancers, multiple sclerosis, inflammatory diseases and HIV. They are dispensed by specialty pharmacies, which must verify benefits, obtain prior authorization, assess clinical appropriateness, arrange financial assistance and schedule delivery before the first fill. The time this takes is a recognized quality measure: the Pharmacy Quality Alliance defines specialty pharmacy turnaround time as the number of days between receipt of a new prescription and the point at which it is ready for pickup or scheduled for delivery [1]. Focus groups of specialty pharmacy staff named prior authorization and third-party issues, patient communication and prescriber issues as the main impediments to turnaround [2], and in a national survey of 1,001 physicians conducted in December 2022, 94% reported that prior authorization delays access to necessary care and 80% that it can lead patients to abandon recommended treatment [3].

Before any of that work can start, the prescription has to be taken in. Intake is the stage at which a referral becomes a case: the patient is identified, demographic and insurance data are captured, the prescriber is verified, and the case is routed to benefits, clinical and fulfillment teams. Intake is also labor-intensive. Pharmacy technicians do most of it, and they are hard to retain: in workforce surveys conducted for the American Society of Health-System Pharmacists in late 2021, a majority of pharmacy administrators reported technician turnover of at least 21% for that year, and nearly nine in ten reported assigning technician tasks to pharmacists [4]. Technicians are also where data-entry errors are caught: in a study of community pharmacies, 92% of observed e-prescription errors were detected during data entry, most of them by technicians [5].

A structural feature of specialty referrals makes intake harder than it needs to be. Many patients start therapy on several medications at once: a combination regimen, a supportive medication alongside a cancer therapy, or the supplies that accompany an injectable. These prescriptions are written together but arrive separately. One e-prescription may arrive

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

seconds before a second; a faxed referral form with the remaining drugs arrives only after it has been scanned and read; and an occasional prescription follows a day later. Intake systems inherited from retail dispensing treat each arriving prescription as the unit of work. For a multi-prescription episode this produces several cases for one patient, each needing its own identification, demographic capture and benefit verification. It creates opportunities for those cases to carry conflicting data, and it pushes reconciliation downstream, where several teams may contact the same patient about the same course of therapy.

Consolidating an episode into one case raises three engineering questions. First, when has an episode arrived? Prescriptions form an unbounded stream with no end-of-episode marker, so every grouping rule trades the completeness of a group against the delay it adds. Second, which documents belong to the same patient? Documents arrive through channels with different identifiers and different error profiles, and a patient's name or date of birth may be recorded differently on each. Exact matching splits episodes, while permissive matching risks merging two patients, which is a patient-safety failure. Third, how should automated clearance work on a consolidated case? Information missing from one document is often present on another, but documents can also disagree.

This paper answers these questions with a vendor-neutral reference architecture and evaluates it by simulation. The architecture generalizes design patterns used in large-scale specialty pharmacy intake; the paper presents them without reference to any particular organization's systems or data. Its contributions are:

- a formulation of multi-prescription intake as keyed event correlation over an unbounded, noisy stream, with metrics for fragmentation, consolidation, added hold, labor and straight-through processing (Section III);
- a reference architecture that combines a federated entity-resolution service, a session-window grouping engine with a bounded hold, an urgency bypass and late attachment, and STP rules evaluated on the consolidated case at the time each decision is actually made (Section IV);
- a probabilistic patient-linking rule with an ambiguity margin, which declines to link automatically when two candidate records score alike, as happens with twins (Section IV-B);
- a simulation study with 30 replications per configuration, five policies, parameter sweeps, ablations, an identity stress test and a sensitivity analysis, with paired non-parametric tests (Sections V and VI).

Section II reviews related work. Section III formulates the problem and Section IV describes the architecture and its algorithms. Section V describes the experimental method and Section VI reports the results. Section VII discusses implications and limitations, and Section VIII concludes.

II. BACKGROUND AND RELATED WORK

A. Specialty pharmacy intake and turnaround

Turnaround time has become a central performance measure for specialty pharmacies [1]. Gabriel et al. [2] held focus groups with specialty pharmacy staff and found prior authorization and third-party payer issues raised as barriers in every group, together with patient-communication and prescriber issues. Prior authorization is also the subject of federal rulemaking: the Interoperability and Prior Authorization final rule (CMS-0057-F) requires impacted payers to support electronic prior authorization interfaces [6] based on HL7 FHIR [7] and the Da Vinci Prior Authorization Support implementation guide [8]. On the prescribing side, e-prescriptions in the United States follow the NCPDP SCRIPT standard [9]. These standards make individual transactions electronic, but they do not say how a pharmacy should assemble the transactions that belong to one patient's course of therapy. That assembly is the subject of this paper.

B. Event correlation and windowing

Grouping related events that arrive over time is a core problem of stream processing. The Dataflow Model [10] distinguishes fixed (tumbling), sliding and session windows and separates what is computed from when results are emitted; session windows, which close after a gap of inactivity, are its model for bursts of user activity. Apache Flink [11] and similar engines implement such windows with durable state and event-time semantics. Complex event processing extends correlation to patterns across event types [12], [13]. In enterprise integration, the Aggregator pattern collects related messages until a completeness condition or a timeout is met [14]. Dataflow also handles data that arrive after a window has fired: an allowed-lateness horizon and accumulation modes decide whether late data update a result already emitted [10]. Late attachment in this paper plays the same role, with a domain-specific horizon, the milestone after which a case's contents are fixed. The trade-off between waiting and acting is old: Nagle's algorithm holds small packets briefly so that they can be sent together [15]. Intake grouping is a session-window aggregation keyed by patient, with two healthcare-specific differences: the key must be inferred from noisy identity data, and some events, those for urgent therapies, must never wait.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

C. Patient matching and record linkage

Probabilistic record linkage originates with Fellegi and Sunter [16], who showed how agreement and disagreement on individual fields can be weighted by likelihood ratios and combined into a decision with link, non-link and review regions. Jaro [17] and Winkler [18] introduced string comparators suited to names. Christen [19] and Binette and Steorts [20] review the field, and Papadakis et al. [21] survey the blocking techniques that make linkage scale. In healthcare, patient matching is a long-standing problem [22]: discrepancies in key identifying fields are common in master patient indexes [23], and standardizing demographic data improves matching accuracy [24]. HL7 FHIR defines a Patient $match operation that returns candidate records with match scores [7], so scored, probabilistic matching is now part of standard interfaces. Learned matchers, from deep networks over attribute embeddings [25] to fine-tuned pre-trained language models [26], have raised accuracy on benchmark matching tasks, and referential matching compares incoming records with a curated reference set of identities rather than only with one another [27]. Any of these can replace the scorer used here: the architecture needs only a score, calibrated thresholds and a margin between the best and second-best candidates. Intake differs from batch record linkage in two ways. Decisions are made online, one document at a time, and the population being matched includes patients that intake itself created minutes earlier.

D. Data federation and entity services

Data-integration theory distinguishes virtual integration, in which queries over a mediated schema are reformulated against the sources, from approaches that materialize integrated data [28], [29]. Microservice architectures commonly expose such integrated views through API composition and query services that sit over several systems of record [30], [31]. The integrated view intake needs is narrow but latency-sensitive: given a document, return the patient it most likely belongs to, drawing both on golden records held by enterprise systems and on provisional records that intake has created but not yet reconciled.

E. Positioning

Work on specialty pharmacy operations identifies the delays but not the intake architecture that would remove the fragmentation behind some of them, and work on stream processing and record linkage supplies components but not their combination under healthcare constraints. This paper combines them and evaluates the combination, with explicit treatment of urgent therapies, late documents and the timing of automated clearance decisions. Table II contrasts the intake models compared in this study, and Table I summarizes which of the required capabilities each line of related work provides.

TABLE I. CAPABILITIES PROVIDED BY RELATED APPROACHES

Approach	Keyed grouping	Uncertain keys	Bounded wait	Urgent bypass	Late items	Decision timing
Stream windows and triggers [10], [11]	yes	no	yes (triggers)	custom trigger	allowed lateness	not addressed
Aggregator pattern [14]	correlation identifier	no	timeout	no	no	not addressed
Complex event processing [12], [13]	yes	no	time windows	possible	limited	not addressed
Probabilistic record linkage [16], [19]	n/a	yes	n/a	n/a	n/a	batch
Patient matching in health IT [22], [24]	n/a	yes	n/a	n/a	n/a	n/a
This paper	resolved patient	yes (margin rule)	cap H	yes	until milestone	at release and attachment

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

TABLE II. INTAKE MODELS COMPARED IN THIS STUDY

Model	Grouping key	Case created	Added hold	Later documents	Urgent prescriptions
Per-prescription	none	on arrival	none	new case	not held
Tumbling window W	patient	at window boundary	up to W	new case	wait for boundary
Session window g	patient	after gap g	g or more, unbounded	new case	wait for gap
Immediate release + attach	resolved patient	on first arrival	none	attached until milestone	not held
Proposed	resolved patient	after gap g or cap H	at most H	attached until milestone	released at once

III. PROBLEM FORMULATION

Each arriving document, here a prescription with its accompanying referral data, is a tuple $d_i = (a_i, x_i, ch_i, u_i, A_i)$, where a_i is the arrival time, x_i the observed patient identity (names, date of birth, ZIP code, phone, external identifier), ch_i the channel, u_i an urgency flag and A_i the administrative and clinical attributes the document carries (completeness of demographics, insurance identifier, diagnosis). Documents are generated by episodes. An episode e is one patient's set of prescriptions for a course of therapy, written together but arriving at different times. Episodes are not observed; only documents are.

An intake policy maps each document to a case $c(i)$ and fixes the release time r_c at which the case is created and work on it begins. Let $C(e)$ be the set of distinct cases holding the documents of episode e, and E_m the set of episodes with two or more prescriptions. Fragmentation F and consolidation completeness Q are

$$F = \frac{1}{|E_m|} \sum_{e \in E_m} |C(e)| \quad (1)$$

$$Q = \frac{1}{|E_m|} \sum_{e \in E_m} \mathbb{1}\left[\, |C(e)| = 1 \,\right] \quad (2)$$

Per-prescription intake has F equal to the mean size of multi-prescription episodes and $Q = 0$; complete consolidation has $F = 1$ and $Q = 1$. The added hold of a document present at release is

$$h_i = r_{c(i)} - a_i \quad (3)$$

and documents attached to an already released case have $h_i = 0$. Intake labor, in minutes, is modeled as

$$L = n_c\, s_c + n_0\, s_r + n_a\, s_a \quad (4)$$

where n_c is the number of cases, n_0 the number of prescriptions present when their case is released, n_a the number attached later, s_c the per-case cost of identification and demographic and benefit verification, s_r the per-prescription cost, and s_a the cost of a late attachment, which reopens a case already in work, adds the drug, repeats its benefit check and contacts the patient again. Labor is reported per 1,000 prescriptions. A prescription is processed straight through when no technician has to touch it; the predicate is defined in Section IV-D. Technician minutes count only work that is not straight through.

Two constraints apply. False merges, that is cases holding documents of two different patients, must not occur in practice, because a merged case can associate the wrong drug or insurance with a patient. Urgent prescriptions must not be held: $h_i = 0$ whenever u_i is set. The design problem is to choose a policy that minimizes labor and maximizes the straight-through share subject to these constraints and to a bound H on the hold of any document. Table III lists the notation.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

TABLE III. NOTATION

Symbol	Meaning
d_i, a_i, x_i	document i, its arrival time, its observed identity
e, E_m, $C(e)$	episode; episodes with two or more prescriptions; cases holding episode e
F, Q	fragmentation (cases per multi-prescription episode); share of such episodes in one case
g, H, M	inactivity gap; hold cap; milestone after which a case accepts no documents
r_c, h_i	release time of case c; added hold of document i
s_c, s_r, s_a	labor per case, per prescription, per late attachment (minutes)
S, τ_{hi}, τ_{lo}, δ	match score; auto-link threshold; review threshold; ambiguity margin

IV. PROPOSED ARCHITECTURE

Fig. 1 shows the architecture. Channel adapters convert each arriving prescription into a canonical intake event. The entity-resolution service links the event to a patient through a federated identity view. The temporal grouping engine holds events in session windows keyed by the resolved patient and releases each window as an episode case. The case assembler pools attributes across the documents in a case and detects conflicts, and embedded STP rules decide which prescriptions can proceed without a technician. A late-attach router adds documents that arrive after release to the open case until the case reaches a milestone after which its contents are fixed. Fig. 2 illustrates the grouping rules on three patients.

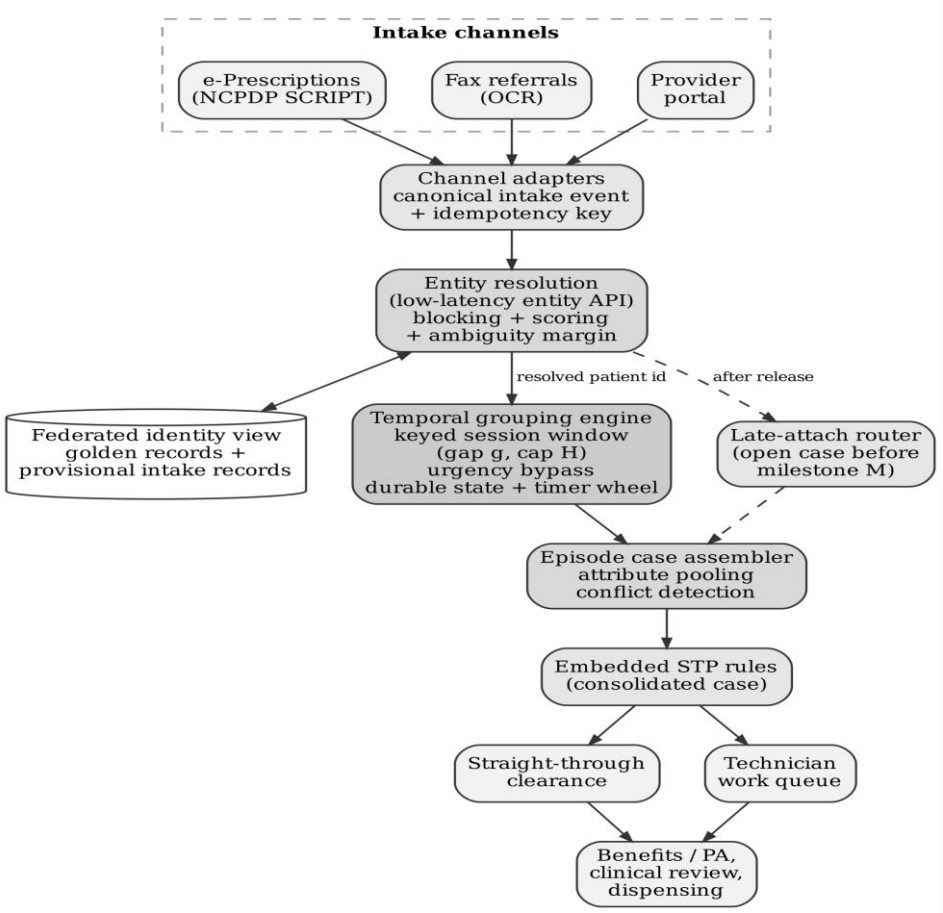


Fig. 1. Reference architecture for patient-centric intake. Dashed arrows show late attachment of documents that arrive after a case was released.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

A. Canonical intake events

Each channel adapter normalizes its input, whether an NCPDP SCRIPT new-prescription message [9], a fax referral read by optical character recognition, or a portal submission, into a canonical event that carries the identity fields, the prescription, the prescriber and whatever benefit and clinical data the source contains, together with the channel and the source timestamp. Each event carries an idempotency key derived from the source message identifier, so that a redelivered message is recognized and processed once [32]. Events are appended to a durable log [33]. The log is the system of record for intake and allows grouping state to be rebuilt by replay.

Table IV lists the main fields of the canonical event. Fields are optional except the envelope: a fax may carry no external identifier and an e-prescription no insurance, and the design must work with whatever each document provides.

TABLE IV. CANONICAL INTAKE EVENT: MAIN FIELD GROUPS

Field group	Examples	Used by
Envelope	event identifier, idempotency key, channel, source and receipt timestamps	log, de-duplication, ordering
Patient identity	names, date of birth, sex, ZIP code, phone, external identifiers such as a medical record or member number	entity resolution
Prescription	drug (NDC or RxNorm code), quantity, directions, days' supply, refills	case assembly, STP
Prescriber	name, NPI, contact details	case assembly, STP
Benefits	payer, BIN/PCN/group, member identifier	case assembly, STP, conflict detection
Clinical	diagnosis (ICD-10-CM), height and weight, allergies when supplied	STP, clinical review
Context	urgency flag, therapy class, referral source	urgency bypass, routing
Attachments	referral images, OCR confidence	review, audit

B. Federated entity resolution

The resolution service answers one question quickly: which patient does this document belong to? Its index federates two populations, golden records held by enterprise systems of record for patients already known to the pharmacy, and provisional records created by intake for new patients and not yet reconciled. Including provisional records is essential. In a new patient's referral episode, the second document must match the record created minutes earlier for the first.

Candidates are generated by blocking [21] on four keys: the exact date of birth; the date of birth with day and month exchanged, a common transcription error; the Soundex code of the last name with the birth year; and the phone number. Each candidate record y is scored against the observed identity x with Fellegi–Sunter field weights [16]

$$w_k(x,y) = \begin{cases} \log \frac{m_k}{u_k} & \text{field } k \text{ agrees} \\ \log \frac{1-m_k}{1-u_k} & \text{field } k \text{ disagrees} \\ 0 & \text{field } k \text{ missing} \end{cases} \quad (5)$$

$$S(x,y) = \sum_k w_k\,(x,y) \quad (6)$$

where m_k and u_k are the probabilities that field k agrees for matching and non-matching pairs. Names are compared with the Jaro–Winkler similarity [17], [18], with agreement thresholds of 0.90 for first names, which also accept common nicknames, and 0.92 for last names; a date of birth that matches after exchanging day and month receives partial credit. The weights used in this study (Table VIII) are illustrative; in practice they are estimated from labeled pairs or by expectation maximization [19]. Let $S^{(1)}$ and $S^{(2)}$ be the best and second-best candidate scores. The decision is

$$\text{decision} = \begin{cases} \text{link to } y^{(1)} & S^{(1)} \geq \tau_{hi} \text{ and } S^{(1)} - S^{(2)} \geq \delta \\ \text{human review} & S^{(1)} \geq \tau_{lo} \text{ otherwise} \\ \text{new provisional record} & \text{otherwise} \end{cases} \quad (7)$$

The margin δ addresses a specific hazard. Twins share a last name, date of birth, address and often a phone number; their records differ in first name and in external identifiers that a given document may not carry. When a document's first name is misspelled or abbreviated, both twins' records can score above the link threshold with nearly equal scores,

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

and a rule that links to the top-scoring record then chooses a sibling almost at random. Requiring a margin between the best and second-best candidates sends exactly these cases to a person.

Algorithm 1 gives the resolution procedure. New provisional records are added to the index immediately, so that the next document of the same new patient finds them.

Algorithm 1. Online entity resolution with an ambiguity margin

```
resolve(x):
  C <- union of index.block(key) for key in blocking_keys(x)
       # exact DOB, swapped DOB, Soundex(last)+birth year, phone
  if C is empty: return new_provisional(x)
  for y in C: s[y] <- sum over fields k of w_k(x, y)
  y1, y2 <- best and second-best candidates by s
  if s[y1] >= tau_hi and s[y1] - s[y2] >= delta: return link(y1)
  if s[y1] >= tau_lo: return review(x, [y1, y2])   # a person decides
  return new_provisional(x)
new_provisional(x):
  r <- provisional record built from x; index.add(r)
  return link(r)
```

C. Temporal grouping algorithm

For each resolved patient key k, the grouping engine keeps at most one open window, with first-arrival time f_k and last-arrival time ℓ_k. The window is released at

$$r_k = \min(\ell_k + g,\ f_k + H) \tag{8}$$

that is, after a gap g with no new document or when the cap H after the first document is reached, whichever comes first. The gap makes the window adaptive: a burst of documents extends it, and a patient whose prescriptions arrive together is released soon after the last one. The cap bounds the hold of every document by H, so a slow trickle of documents cannot hold a case indefinitely. Two further rules handle healthcare constraints. When a document with the urgency flag arrives, its window is released at once with whatever it holds (urgency bypass). And a document that arrives after its patient's window was released is attached to the released case c if that case has not yet reached its milestone M_c, the point after which its contents are fixed, such as the start of clinical review or of the patient scheduling call:

$$\text{attach } d_i \text{ to } c \Leftrightarrow r_c < a_i \le r_c + M_c \tag{9}$$

Algorithm 2 summarizes the procedure. Window state is stored durably and keyed by patient. Timers are kept in a hashed timing wheel [34], so starting, resetting and firing a timer each cost $O(1)$.

Algorithm 2. Patient-keyed session grouping with bounded hold, urgency bypass and late attachment

```
on_event(d):
  k <- resolve(d.identity)            # Section IV-B
  if open window w exists for k:
    w.add(d); w.last <- d.t
    reset_timer(w, min(w.last + g, w.first + H))
  else if released case c of k and c.release < d.t <= c.milestone:
    attach(c, d); decide_stp(c, [d]); return
  else:
    w <- new_window(k, first = d.t, last = d.t); w.add(d)
    set_timer(w, d.t + min(g, H))
  if d.urgent: release(w)             # urgency bypass
on_timer(w): release(w)
release(w):
  c <- assemble_case(w.documents)     # pool attributes, flag conflicts
  decide_stp(c, w.documents)          # decisions made at release
  persist(c); cancel_timer(w); close(w)
```

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

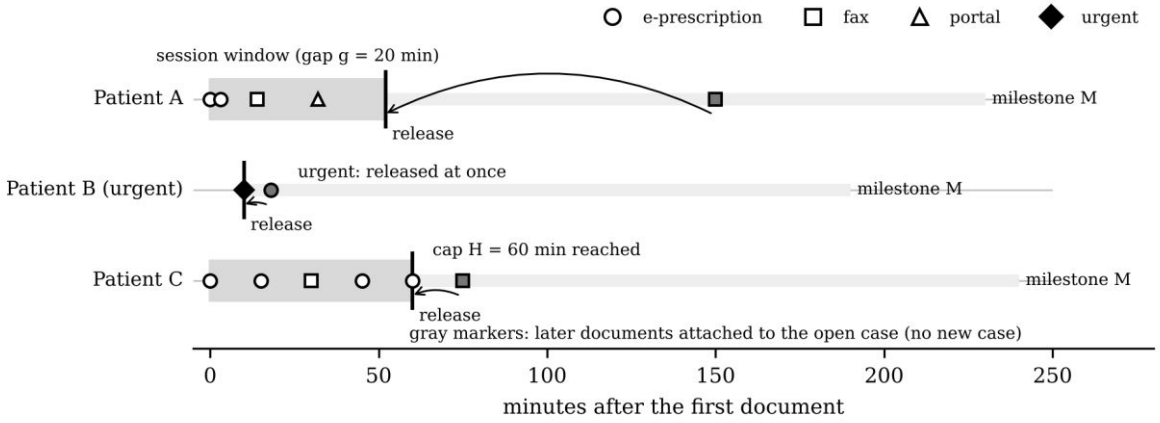


Fig. 2. Grouping rules on three patients: a burst released after the gap (A), an urgent prescription released at once (B) and a trickle closed by the cap (C). Gray markers are later documents attached to the released case.

D. Case assembly and the timing of STP decisions

When a window is released, the case assembler pools the attributes of its documents: demographics that are complete on any document, insurance identifiers and diagnoses. If documents carry different insurance identifiers, the conflict is flagged and the case goes to a technician. Under per-prescription intake the same conflict goes unnoticed, because each case holds only one document. The straight-through predicate for prescription i in case c at time t is

$$\mathrm{STP}(i,t) = D(c_t) \wedge I(c_t) \wedge X(c_t) \wedge \neg f_i \wedge \neg p_i \qquad (10)$$

where D, I and X state that the documents in c at time t provide complete demographics, exactly one insurance identifier and a diagnosis; f_i marks a clinical flag requiring pharmacist review; and p_i marks a prior authorization that is required and not yet on file. When the predicate is evaluated matters as much as what it says. It is evaluated at release for the documents present then, and at attachment for each late document. A late document can benefit from the information already in the case, but it cannot change decisions already made: once a case has been sent to a technician because insurance was missing, the technician's work has been done even if the next document supplies the insurance.

E. Worked example

Table V traces a hypothetical episode through the design. A patient new to the pharmacy is referred with two e-prescriptions sent within a minute and a fax, read by optical character recognition, that carries a third prescription and the insurance card. A fourth prescription arrives the next morning, after the scheduling call that marks the case milestone has taken place.

TABLE V. TRACE OF A HYPOTHETICAL EPISODE (20-MINUTE GAP, 60-MINUTE CAP)

Time	Event	Decision
09:02:10	e-prescription 1 (drug A)	no candidate scores above τ_{lo}: provisional record P created; window opened, release due 09:22:10
09:02:41	e-prescription 2 (drug B)	links to P with a clear margin; release due 09:22:41
09:19:05	fax: drug C and insurance card; first name shortened by OCR	links to P on last name, date of birth and phone; release due 09:39:05
09:39:05	gap expires	one case with three prescriptions; insurance from the fax applies to all three; drug A cleared straight through, drug B needs prior authorization, drug C flagged for pharmacist review
11:30	scheduling call made	milestone reached: the case accepts no further documents
next day 08:40	e-prescription 4 (supportive drug)	links to P; milestone passed, so a new case is opened for the same patient

F. Correctness and fault tolerance

Three properties matter in production. The idempotency key prevents a redelivered document from creating a second case [32]. Windows and timers are checkpointed with the log offset they reflect, so after a failure the engine resumes

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

from the last checkpoint and replays later events, the standard approach in stateful stream processors [11]. And if a reviewer later finds that a case holds two patients' documents, the case is split by a compensating transaction rather than edited in place, preserving the audit trail [35]. Every decision, whether a link, a referral to review, a release, an attachment or an STP outcome, is emitted as an event, so the reasoning behind every case can be reconstructed.

G. Implementation on stream processors

The grouping engine maps onto the keyed state and timers of common stream processors. In Apache Flink, a keyed process function holds the window for each patient and registers a timer for $\min(\ell + g, f + H)$, resetting it on every new document [11]. Kafka Streams offers session windows with an inactivity gap and a grace period for late records [36]; the cap and the urgency bypass, which a plain session window does not provide, require a custom processor over the same state stores. Processing time is the right clock for the hold, which is a promise about wall-clock delay to patients, while source timestamps are kept for ordering and audit [10]. The resolution service is a stateless API over a replicated index. Its tail latency matters more than its mean, because every document waits for it, and hedged or replicated requests are the standard remedy [37].

H. Complexity

Table VI gives the per-document cost of each component. Resolution dominates; its cost is the number of blocked candidates times the number of compared fields. In the simulation, a single-threaded Python implementation resolved about 11,770 documents per second, with a mean of 10.2 candidate comparisons per document against an index of 60,000 golden records. That rate is far above the intake rate of a large pharmacy, and the other components cost constant time per document.

TABLE VI. PER-DOCUMENT COST OF EACH COMPONENT

Component	Time per document	State
Channel adapter	O(message size)	none
Entity resolution	$O(B \cdot K)$, B blocked candidates, K fields	index over N records, $O(N)$
Grouping engine	$O(1)$ amortized (hash map, timing wheel)	O(open windows)
Case assembly and STP	O(documents in case)	O(open cases)

V. EXPERIMENTAL METHOD

A. Simulation design

A discrete-event simulation was written in Python. Each replication generates eight operating days of referrals and discards the first day as warm-up [38]. Referrals arrive at 1,200 episodes per 10-hour operating day; with the episode-size distribution in Table VIII this gives about 2,520 prescriptions per operating day and about 17,736 measured prescriptions per replication. Within an episode, the first document arrives at the episode start. Each further document arrives on the same channel shortly afterwards (exponential delay, mean 2 minutes, probability 0.55), on another channel after a scanning or portal delay (log-normal, median 12 minutes, $\sigma = 1.0$), or, with probability 0.08, as a straggler 2 to 30 operating hours later. Channels are e-prescription (60%), fax (30%) and portal (10%). Forty percent of episodes belong to patients new to the pharmacy; the rest belong to one of 60,000 synthetic patients in the golden index, of whom 1.5% are twins. Five percent of episodes begin with an urgent prescription. Each document carries the patient's identity perturbed by a channel-specific noise model and administrative attributes drawn with channel-specific completeness (Table VIII). A released case reaches its milestone after a log-normal time with median 3 hours.

B. Policies

Five policies are compared on identical document streams: per-prescription intake; a tumbling window of 30 minutes; a session window with a 20-minute gap and no cap; immediate release with late attachment, which is the proposed design with $g = 0$; and the proposed design with $g = 20$ minutes and $H = 60$ minutes. All grouping policies use the same resolved patient identifiers, so differences among them are due to grouping alone. Sweeps vary g from 0 to 90 minutes, the tumbling width from 5 to 90 minutes and H over 30, 60 and 120 minutes. Ablations remove the cap, the urgency bypass, late attachment, or federated resolution, which is replaced by an exact key of last name, first name and date of birth.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

C. Labor and staffing

Labor follows Eq. (4) with s_c = 7, s_r = 3 and s_a = 6 minutes. These unit times are assumed, not measured, and Section VI-H shows how the results change when they vary. Technician minutes count, for each case, s_c plus s_r for each prescription that is not straight through at release, and s_a for each late attachment that is not straight through. To express minutes as staff, 510 productive minutes per technician per 10-hour day (85%) are assumed. Separately, a queueing simulation of a technician pool, first come first served with priority for urgent cases, finds the smallest pool for which the 95th-percentile wait stays within 60 minutes over the measured week; by Little's law [39] this pool tracks the offered load, and it corresponds to operating close to full utilization [40].

D. Identity stress test

Because twins rarely both appear in a week of referrals, wrong links are evaluated separately. For each of six seeds, a population with 5% twins is generated; 1,500 twin pairs and 1,500 randomly chosen patients are selected; and each patient's identity is observed three times through randomly chosen channels. Each observation is resolved against the golden index for auto-link thresholds from 6 to 14, with and without the ambiguity margin δ = 3 and without a review band. Outcomes are a correct link, a wrong link, or no automatic link.

E. Verification of the grouping implementation

The session-window implementation was checked against a closed form. If the offsets of the $K-1$ later documents of an episode, measured from the first, are independent exponential variables with rate μ, the spacings between consecutive arrivals are independent exponential variables with rates $(K-1)\mu$, $(K-2)\mu$, …, μ (the Rényi representation of exponential order statistics [41]). A session window splits the episode at every spacing longer than g, so

$$\mathbb{E}[\,|C(e)|\,] = 1 + \sum_{m=1}^{K-1} e^{-m\mu g} \qquad (11)$$

Table VII compares Eq. (11) with the implementation on 20,000 synthetic episodes per cell, with mean offset $1/\mu$ = 10 minutes. The largest absolute difference is 0.010 cases per episode, within Monte Carlo error.

TABLE VII. SESSION-WINDOW FRAGMENTATION: CLOSED FORM AGAINST IMPLEMENTATION (MEAN OFFSET 10 MIN)

K	g = 5	g = 10	g = 20	g = 40
2	1.607 / 1.605	1.368 / 1.371	1.135 / 1.137	1.018 / 1.018
3	1.974 / 1.972	1.503 / 1.493	1.154 / 1.151	1.019 / 1.017
4	2.198 / 2.195	1.553 / 1.562	1.156 / 1.151	1.019 / 1.017
6	2.415 / 2.418	1.578 / 1.586	1.157 / 1.155	1.019 / 1.017

Each cell shows the closed form followed by the simulated value.

F. Statistics and environment

Each configuration is run for 30 independent replications with common random numbers across policies. Means are reported with 95% confidence intervals from the t distribution. Policies are compared with two-sided Wilcoxon signed-rank tests paired by replication [42], with Holm correction for multiple comparisons [43]. Experiments ran on an x86-64 Linux workstation (8-core CPU, 32 GB of memory) with Python 3.11.4, NumPy 1.24.3 and SciPy 1.10.1. Table VIII lists the baseline parameters. They are assumed inputs chosen to be plausible, not measurements from any pharmacy or production system, and Section VI-H varies them.

TABLE VIII. BASELINE SIMULATION PARAMETERS

Parameter	Value
Episodes per operating day; days per replication	1,200; 8 (1 warm-up)
Episode size K = 1, …, 6	0.45, 0.25, 0.15, 0.08, 0.04, 0.03
Channel mix (e-prescription, fax, portal)	0.60, 0.30, 0.10
Within-episode delays	same channel Exp(mean 2 min), p = 0.55; other channel LogN(median 12 min, σ = 1.0)
Stragglers	probability 0.08, uniform 2–30 operating hours
New patients; urgent episodes	40%; 5%
Milestone after release	LogN(median 3 h, σ = 0.5)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

Parameter	Value
Complete demographics (eRx, fax, portal)	0.97, 0.80, 0.95
Insurance identifier present	0.55, 0.45, 0.90 (outdated: 0.02, 0.05, 0.01)
Diagnosis present	0.85, 0.75, 0.90
Clinical flag; prior authorization needed	0.30; 0.30 per prescription
Name typo probability (eRx, fax, portal)	0.02, 0.08, 0.01 per name field
Missing ZIP, phone, external ID (fax)	0.15, 0.30, 0.80
Field weights agree/disagree	first 3.0/−2.5; last 4.0/−3.5; DOB 5.0/−6.0 (swapped +2.0); ZIP 1.5/−1.0; phone 3.0/−0.5; external ID 8.0/−8.0
τ_{hi}, τ_{lo}, δ	11, 6, 3
s_c, s_r, s_a (minutes)	7, 3, 6

VI. RESULTS

A. Overall comparison

Table IX summarizes the five policies; confidence intervals are narrow (for example, ±9 minutes per 1,000 prescriptions for the labor of the proposed design). Per-prescription intake creates one case per prescription, so a multi-prescription episode is split into 3.00 cases on average. All grouping policies reduce fragmentation, but in different ways. The tumbling window releases at fixed boundaries and splits episodes that straddle a boundary, consolidating 53.5% of multi-prescription episodes. The session window adapts to bursts and consolidates 68.8%, but documents arriving after its window closes open new cases. Late attachment closes most of that gap: both attachment policies consolidate about 85% of episodes, and their remaining fragmentation comes from stragglers that arrive after their case passed its milestone.

TABLE IX. MAIN RESULTS, MEAN OF 30 REPLICATIONS

Policy	Cases / 1,000 Rx	*F*	*Q* (%)	p95 hold (min)	Urgent p95 hold (min)	Intake min / 1,000 Rx	Technician min / 1,000 Rx	STP (%)
Per-prescription	1,000	3.00	0.0	0.0	0.0	10,000	7,887	21.1
Tumbling window (W = 30)	621	1.54	53.5	28.5	28.5	7,346	5,559	32.8
Session window (g = 20)	571	1.35	68.8	36.3	36.4	6,996	5,259	34.3
Immediate release + attach (g = 0)	519	1.16	85.1	0.0	0.0	8,078	5,834	30.0
Proposed (g = 20, H = 60)	517	1.15	85.5	36.0	0.0	6,857	5,070	35.0

The difference between immediate release and the proposed design shows what the hold buys. With $g = 0$, 48% of prescriptions arrive after their case has been released and must be attached, each time reopening a case already in work; the proposed design attaches only 7.8%. The proposed design therefore needs 15.1% fewer intake minutes than immediate release, and it processes 5.1 percentage points more prescriptions straight through, because STP decisions for the first prescriptions of a burst are made with information from the whole burst.

Against per-prescription intake, the proposed design creates 48.3% fewer cases, needs 31.4% fewer intake minutes and 35.7% fewer technician minutes, and raises the straight-through share from 21.1% to 35.0%. In staffing terms, technician work falls from 15.5 to 9.9 full-time equivalents per 1,000 daily prescriptions. The queueing simulation gives a consistent picture: the smallest technician pool that keeps the 95th-percentile wait within an hour falls by 35.2%, from 33.4 to 21.6 technicians at the simulated volume.

B. The window trade-off

Fig. 3 shows consolidation as a function of window length. For the tumbling and session windows, consolidation rises steeply with window length and is still incomplete at 90 minutes, because the window must stay open long enough for

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

the slowest documents. For the proposed design, consolidation is almost independent of g, because late attachment catches what the window misses. What g changes is labor (Fig. 4). Intake minutes per 1,000 prescriptions fall from 8,078 at $g = 0$ to 7,012 at $g = 10$, 6,857 at $g = 20$ and 6,749 at $g = 45$ minutes, with diminishing returns beyond about 20 minutes. The session window without attachment needs a gap of 60 minutes (6,707 minutes per 1,000 prescriptions, 95th-percentile hold 93 minutes) to come close. The knee of the proposed curve lies between 10 and 20 minutes, which is why $g = 20$ minutes is the reference setting.

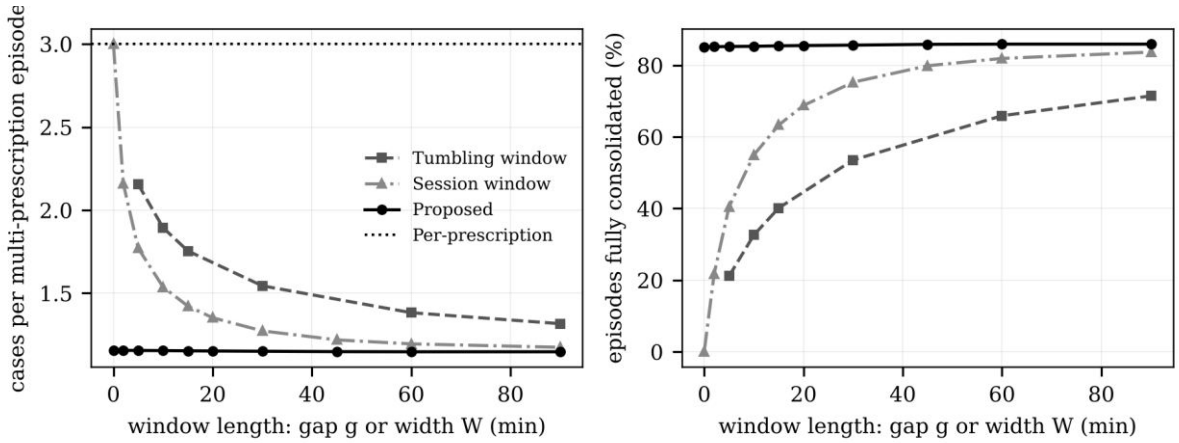


Fig. 3. Fragmentation (left) and share of multi-prescription episodes held in one case (right) as a function of window length. Means of 30 replications; 95% confidence intervals are smaller than the markers.

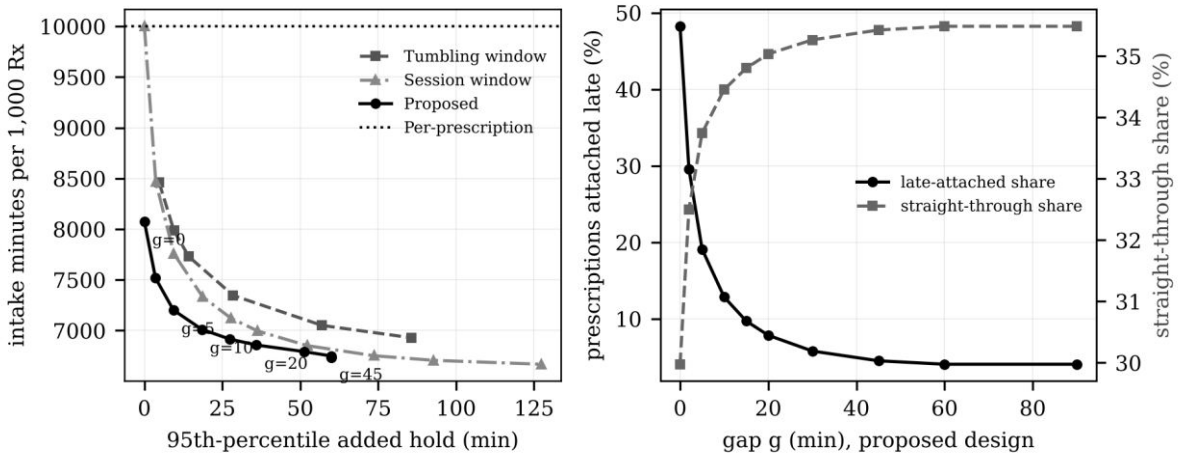


Fig. 4. Left: intake labor against the 95th-percentile added hold as the gap g (session, proposed) or width W (tumbling) varies. Right: share of prescriptions attached late and straight-through share as functions of g in the proposed design.

Table X varies the cap. At $g = 20$ minutes, a cap of 30 minutes bounds every hold at 30 minutes and costs 6 intake minutes per 1,000 prescriptions relative to a cap of 60 minutes, because more documents of slow bursts are attached late. Raising the cap to 120 minutes leaves labor essentially unchanged (6,857 against 6,857 minutes per 1,000 prescriptions), while the worst-case hold rises from 60 to 68 minutes, so caps beyond an hour buy little.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

TABLE X. EFFECT OF THE CAP AND THE GAP (PROPOSED DESIGN, MEAN OF 30 REPLICATIONS)

H (min)	*g* (min)	p95 hold (min)	Max hold (min)	*Q* (%)	Late-attached (%)	Intake min / 1,000 Rx
30	10	18.4	30.0	85.3	12.8	7,012
30	20	30.0	30.0	85.5	8.0	6,863
30	30	30.0	30.0	85.5	6.5	6,817
60	10	18.5	35.6	85.3	12.8	7,012
60	20	36.0	59.9	85.5	7.8	6,857
60	30	51.3	60.0	85.6	5.8	6,792
120	10	18.5	35.6	85.3	12.8	7,012
120	20	36.0	68.0	85.5	7.8	6,857
120	30	51.8	94.2	85.6	5.8	6,792

Choosing g is a policy decision: it trades technician time against waiting time for patients. If an organization is willing to spend λ_h minutes of technician time to save one minute of hold for one prescription, the gap should minimize

$$J(g) = L(g) + \lambda_h \sum_i h_i(g) \qquad (12)$$

where $L(g)$ is the labor of Eq. (4) and the sum runs over all prescriptions. Table XI evaluates Eq. (12) on the simulated sweep. When hold time is treated as nearly free, the largest gaps win; as λ_h grows the optimum moves to shorter gaps, and immediate release with attachment becomes optimal only when a minute of hold is valued at more than a technician-minute. Organizations can make this trade-off explicit instead of setting g by habit.

TABLE XI. GAP THAT MINIMIZES LABOR PLUS A HOLD PENALTY (PROPOSED DESIGN, 60-MINUTE CAP)

λ_h (technician-min per prescription-min)	Optimal *g* (min)	Intake min / 1,000 Rx	Mean hold (min)	p95 hold (min)
0.00	60	6,734	52.7	60.0
0.05	5	7,202	4.7	9.3
0.10	5	7,202	4.7	9.3
0.20	2	7,518	1.6	3.6
0.50	0	8,078	0.0	0.0
1.00	0	8,078	0.0	0.0
2.00	0	8,078	0.0	0.0

C. Labor, straight-through processing and conflicts

Fig. 5 compares labor and STP. Consolidation reduces labor in two ways: one case setup replaces several, and pooled attributes let more prescriptions satisfy the STP predicate, as when a fax without an insurance identifier is cleared because an e-prescription for the same patient carries one. Consolidation also surfaces conflicts. Under the proposed design, 7.7 cases per 1,000 prescriptions are flagged because their documents carry different insurance identifiers, in this model mostly outdated insurance on faxed referrals. Per-prescription intake flags none, not because the conflicts are absent but because each case sees only one document.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

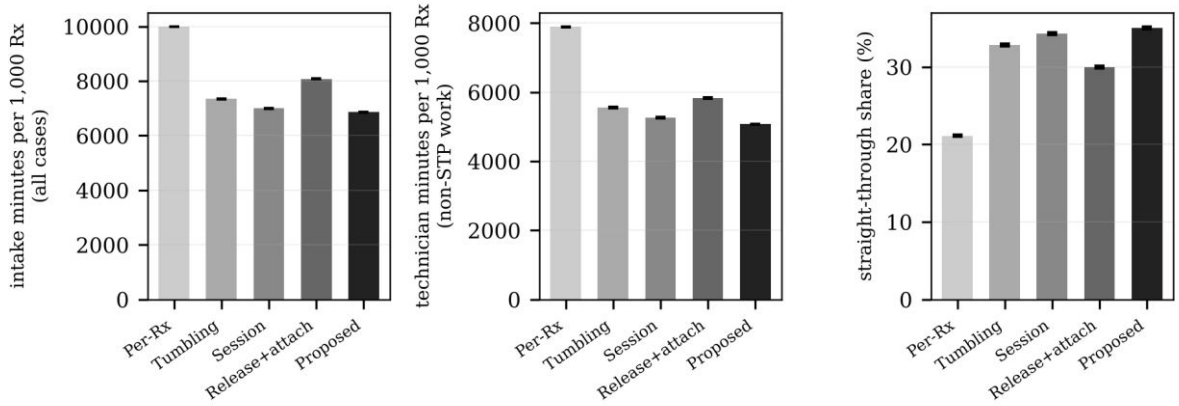


Fig. 5. Intake minutes (all cases), technician minutes (work that is not straight through) and straight-through share by policy. Error bars are 95% confidence intervals.

Fig. 6 turns technician minutes into waiting time. It plots the 95th-percentile wait for a technician against the size of the technician pool, averaged over three replications. Per-prescription intake needs about 33 technicians to keep that wait within an hour at the simulated volume; the proposed design needs about 22. The tumbling window shows a floor of about 8 minutes that more staff do not remove, because all of its cases are released at window boundaries and reach the queue in bursts; session-based policies release cases continuously.

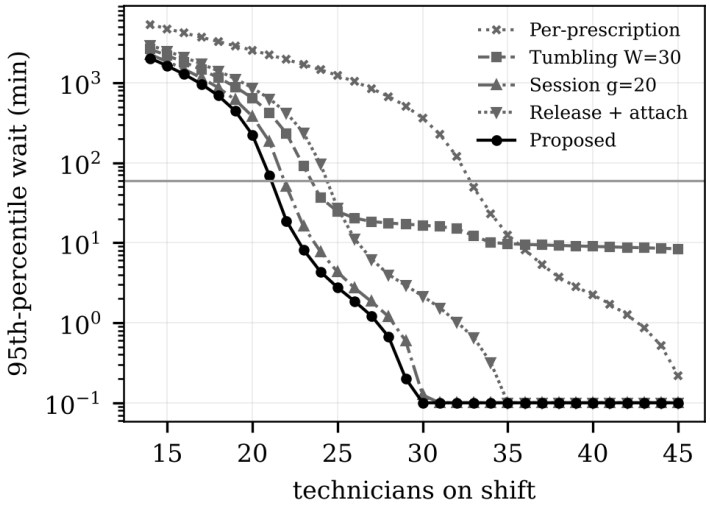


Fig. 6. 95th-percentile wait for a technician against pool size (mean of three replications; waits of zero are plotted at 0.1 minute). The horizontal line marks one hour.

D. Hold latency

Fig. 7 reports the added hold. Under the proposed design the median hold equals g, because single-prescription episodes wait for the gap to expire. The 95th percentile is 36.0 minutes, the 99th 43.8 minutes and the maximum 59.9 minutes, bounded by the cap. Without the cap the maximum rises to 68.0 minutes (Table XIV); at $g = 20$ the cap rarely binds, but it bounds the worst case. Urgent prescriptions are released on arrival instead of waiting 36.4 minutes at the 95th percentile without the bypass. A hold of tens of minutes is small next to specialty turnaround times measured in days [1], and it is spent before any technician work begins.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

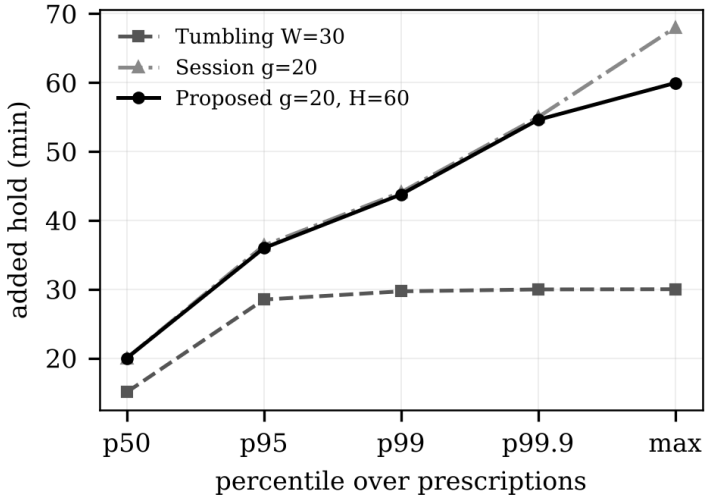


Fig. 7. Added hold before case creation, by percentile over prescriptions. Per-prescription intake and immediate release add no hold.

E. Episode size and channel

Table XII breaks the comparison down by the number of prescriptions in an episode (10 replications). Single-prescription episodes, 46% of all episodes, gain nothing from consolidation: every policy creates one case for them, and their straight-through share is about 21% under every policy because there is nothing to pool. They do pay the hold, a mean of 18.9 minutes under the proposed design, slightly less than g because urgent prescriptions are released at once. The benefit is concentrated in the larger episodes. For an episode of six prescriptions, per-prescription intake creates six cases, the session window 1.73 and the proposed design 1.37, and the straight-through share rises from 21.4% to 43.2% (Fig. 8, left). The mean hold grows only slowly with episode size, from 18.9 to 24.8 minutes, because each new document extends the session until the cap is reached. A pharmacy whose referrals are mostly single prescriptions should therefore expect smaller gains, as the sensitivity analysis also shows, and could use a shorter gap for therapy classes that are rarely prescribed in combination.

TABLE XII. RESULTS BY EPISODE SIZE (10 REPLICATIONS)

Rx in episode	Share of episodes (%)	Cases: per-Rx	Cases: tumbling	Cases: session	Cases: release + attach	Cases: proposed	Mean hold, proposed (min)	STP, proposed (%)
1	45.8	1	1.00	1.00	1.00	1.00	18.9	20.9
2	24.7	2	1.32	1.21	1.08	1.08	19.5	34.1
3	14.8	3	1.57	1.37	1.16	1.16	20.8	39.0
4	7.8	4	1.78	1.50	1.23	1.22	22.1	41.6
5	3.9	5	1.97	1.62	1.31	1.30	23.5	43.0
6	2.9	6	2.15	1.73	1.38	1.37	24.8	43.2

Fig. 8 (right) shows the straight-through share by channel. Faxed documents benefit most: their share rises from 13.3% under per-prescription intake to 30.1% under the proposed design, against 22.2% to 35.9% for e-prescriptions and 37.6% to 44.0% for portal submissions. Faxes are the least complete documents in the model (Table VIII), so they gain most from attributes carried by other documents of the same episode. Immediate release with attachment helps faxes least among the grouping policies (23.8%), because under immediate release the first document of every episode is judged alone, so a fax that opens an episode gains nothing from the documents that follow it minutes later. This is the pooling mechanism behind the overall result, and it indicates where the design should help most in practice: in pharmacies that still receive a large share of referrals by fax.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

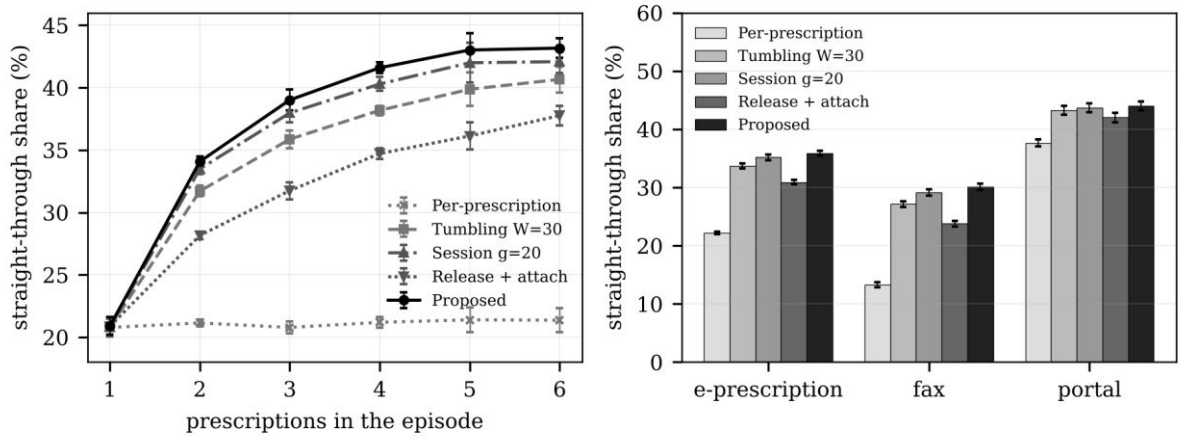


Fig. 8. Straight-through share by number of prescriptions in the episode (left) and by channel (right). Means and 95% confidence intervals of 10 replications.

F. Identity resolution

With an exact key of last name, first name and date of birth, 20.1% of multi-prescription episodes were split across two or more patient keys, because a single typographical error, nickname or exchanged day and month breaks the key. The federated scorer split 0.17%, sent 0.82% of documents to review and produced no false merges in the referral stream (Table XIII). Fig. 9 (left) shows the trade-off as the auto-link threshold varies without a review band: splits rise from 0.8% at a threshold of 7 to 9.3% at 13 and above 40% from 14, while false merges appear only at the lowest threshold of 6 (0.17 per 10,000 documents). The stress test (Fig. 9, right) isolates the twin hazard. Without the ambiguity margin, between 0.04% and 0.23% of twin documents were linked to the wrong sibling, depending on the threshold. With $\delta = 3$, no wrong links occurred in any of the 81,000 resolutions per threshold, at the cost of sending 1.1% of twin documents to review at a threshold of 11. Randomly chosen patients were linked wrongly in at most 0.015% of observations without the margin, and never with it.

TABLE XIII. IDENTITY RESOLUTION IN THE REFERRAL STREAM

Method	Episodes split (%)	False merges	Documents to review (%)
Exact key (last, first name, DOB)	20.1	0	0
Federated scorer ($\tau_{hi} = 11$, $\tau_{lo} = 6$, $\delta = 3$)	0.17	0	0.82

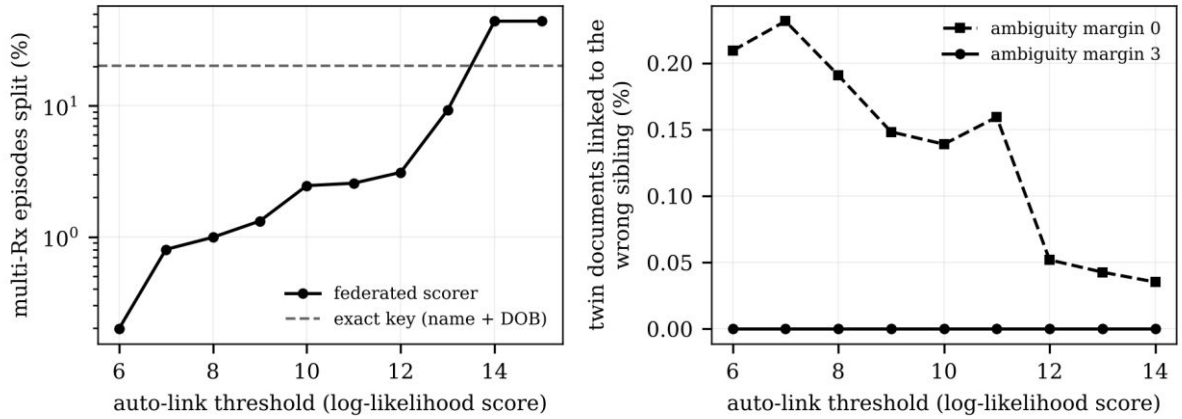


Fig. 9. Left: multi-prescription episodes split across patient keys as the auto-link threshold varies (no review band; log scale). Right: twin documents linked to the wrong sibling in the stress test, without and with the ambiguity margin.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

G. Ablation

Table XIV removes one component at a time. Removing late attachment has the largest effect: fragmentation rises from 1.15 to 1.40 cases per episode and intake minutes rise by 3.2%. Replacing federated resolution with the exact key raises fragmentation to 1.36 and intake minutes by 5.1%. Removing the cap leaves the averages unchanged but raises the maximum hold from 59.9 to 68.0 minutes. Removing the bypass leaves the averages unchanged but makes urgent prescriptions wait 36.4 minutes at the 95th percentile. Each component therefore addresses a distinct failure: late attachment and resolution address fragmentation, while the cap and the bypass address delay.

TABLE XIV. ABLATION OF THE PROPOSED DESIGN (20-MINUTE GAP, 60-MINUTE CAP), MEAN OF 30 REPLICATIONS

Configuration	*F*	*Q* (%)	Max hold (min)	Urgent p95 (min)	Intake min / 1,000 Rx	STP (%)
Proposed (all components)	1.15	85.5	59.9	0.0	6,857	35.0
Without cap *H*	1.15	85.5	68.0	0.0	6,857	35.0
Without urgency bypass	1.15	85.5	59.9	36.4	6,794	35.3
Without late attachment	1.40	65.4	59.9	0.0	7,076	33.9
Exact key instead of federated resolution	1.36	68.9	59.7	0.0	7,208	33.4

H. Sensitivity

Table XV repeats the main comparison under altered assumptions, ten replications each. The reduction in technician minutes of the proposed design relative to per-prescription intake ranges from 29.7% to 40.7%. It is smallest when most episodes have a single prescription and largest when multi-prescription episodes are common, as expected. More stragglers and slower cross-channel arrival shift work from the window to late attachment but leave the ranking of policies unchanged, and doubling the volume leaves per-prescription results unchanged because grouping is per patient. A higher attachment cost penalizes immediate release more than the proposed design, and the case setup cost scales the benefit of consolidation directly (Fig. 10).

TABLE XV. SENSITIVITY OF THE MAIN COMPARISON (10 REPLICATIONS PER SCENARIO)

Scenario	Technician min saved vs per-Rx (%)	Intake min saved vs immediate release (%)	STP per-Rx (%)	STP proposed (%)	p95 hold (min)
Reference assumptions	35.6	15.0	21.2	35.0	36.0
Fewer multi-prescription episodes (55% single)	29.7	12.0	21.1	32.6	34.7
More multi-prescription episodes (30% single)	40.7	17.7	21.2	37.1	36.9
Twice as many stragglers (0.16)	32.7	13.4	21.0	33.8	35.0
Slower cross-channel arrival (median 25 min)	34.1	13.0	21.2	34.4	36.0
Fax-heavy channel mix (50% fax)	36.0	15.0	19.3	33.8	35.9
Twice the volume	35.8	15.1	21.1	35.1	36.1
Cheaper attachment ($s_a = 4$)	36.8	5.8	21.2	35.0	36.0
Costlier attachment ($s_a = 9$)	33.8	25.4	21.2	35.0	36.0
Lower case cost ($s_c = 4$)	31.2	18.6	21.2	35.0	36.0
Higher case cost ($s_c = 10$)	37.9	12.6	21.2	35.0	36.0

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

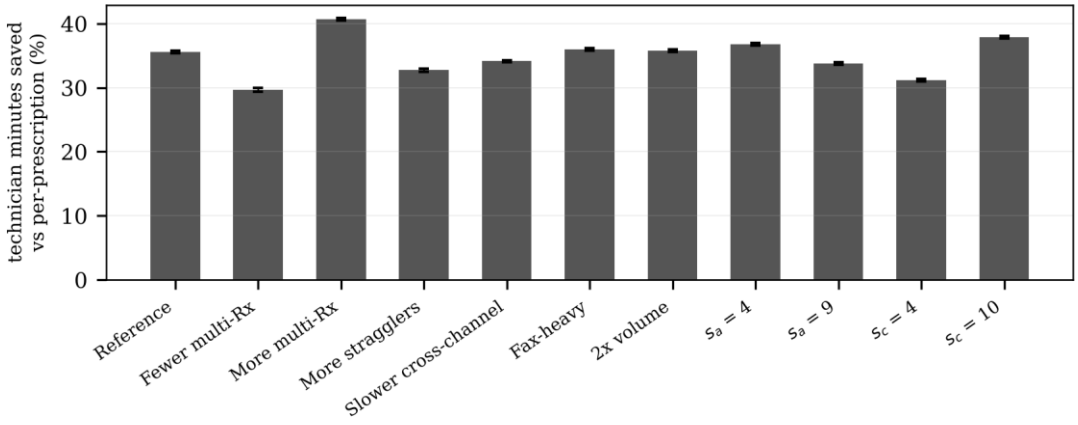


Fig. 10. Reduction in technician minutes relative to per-prescription intake under each sensitivity scenario (mean and 95% confidence interval of 10 replications).

A. Statistical significance

Table XVI reports the paired tests. Every difference between the proposed design and a baseline in the table is significant after Holm correction (largest adjusted p = 1.5×10^{-8}). With 30 replications and differences of the same sign in every replication, the tests mainly confirm that the differences are systematic; the confidence intervals and the sensitivity analysis are the more informative summaries of their size.

TABLE XVI. PAIRED WILCOXON SIGNED-RANK TESTS (30 REPLICATIONS, TWO-SIDED, HOLM-ADJUSTED)

Comparison	Metric	Mean difference	p	Adjusted p
Proposed vs Per-prescription	Intake minutes per 1,000 Rx	-3,142.80	1.9×10^{-9}	1.5×10^{-8}
Proposed vs Per-prescription	Straight-through share	+13.93 pp	1.9×10^{-9}	1.5×10^{-8}
Proposed vs Session window	Cases per multi-Rx episode	-0.20	1.9×10^{-9}	1.5×10^{-8}
Proposed vs Session window	Intake minutes per 1,000 Rx	-138.50	1.9×10^{-9}	1.5×10^{-8}
Proposed vs Session window	Technician minutes per 1,000 Rx	-189.21	1.9×10^{-9}	1.5×10^{-8}
Proposed vs Session window	Straight-through share	+0.76 pp	1.9×10^{-9}	1.5×10^{-8}
Proposed vs Immediate release + attach	Intake minutes per 1,000 Rx	-1,220.73	1.9×10^{-9}	1.5×10^{-8}
Proposed vs Immediate release + attach	Straight-through share	+5.06 pp	1.9×10^{-9}	1.5×10^{-8}

VII. DISCUSSION

A. Operational implications

Two results matter most for practice. First, consolidation is best achieved by combining a short hold with late attachment rather than by a long window. A 20-minute gap with attachment consolidates more episodes than a session window three times as long, and attachment is what allows the hold to be short. Second, the timing of automated decisions matters as much as the grouping itself. Making STP decisions on a consolidated burst rather than on the first document raised the straight-through share by 5.1 percentage points over immediate release in this model. Both results follow from the structure of referral arrivals, bursts followed by occasional stragglers, and should hold wherever that structure holds, but their magnitudes depend on the parameters.

For staffing, the reduction in technician work is the main benefit. With technician turnover high and pharmacists absorbing technician tasks [4], reducing repeated case setup and verification is a lever that does not depend on hiring. It also concentrates human attention where it is needed: in the proposed design, technicians see conflicts and incomplete episodes rather than re-entering the same patient several times.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

B. Patient safety

Consolidation introduces one hazard, merging two patients' documents, and reduces another, conflicting records for one patient. The architecture treats the first as a hard constraint: linking requires both a high score and a clear margin over the next candidate, and ambiguous documents go to a person. The stress test shows that the margin, not the threshold, prevents wrong links between twins; a threshold high enough to prevent them on its own would split many genuine episodes. Compensating transactions allow a mistaken merge found later to be undone with a full audit trail. The second effect is a benefit: conflicting insurance identifiers are detected at intake instead of later in the process.

C. Interoperability

The architecture fits standards-based intake. SCRIPT messages [9] and FHIR-based prior authorization exchanges [7], [8] are further channels into the grouping engine. As the electronic prior authorization interfaces required by CMS-0057-F [6] come into use, the consolidated case is a natural unit for initiating authorizations: one request per episode instead of one per prescription, with the full set of drugs in view.

D. Privacy and security

Intake handles protected health information at every step. The design keeps identity data inside the resolution service and passes only resolved identifiers, and the fields each rule needs, to downstream components, in line with the minimum-necessary standard [44]. The log and the index are encrypted at rest and in transit. Review screens that show candidate records side by side reveal information about other patients and should be limited to trained staff. The decision log, which records who or what made each link, release and STP decision, supports the audit controls required by the HIPAA Security Rule [45].

E. Deployment and monitoring

A deployment should begin in shadow mode. The grouping engine runs alongside the existing intake path, and its decisions are logged but not acted on, so that hold times, attachment rates and match decisions can be compared with what technicians actually do before any case is created by the new path. Weights and thresholds should be fitted to local data and re-validated periodically, and every link decision should be explainable from the field agreements that produced it. Access to the identity index should follow the minimum-necessary principle, and the decision log doubles as an audit trail for protected health information. Table XVII lists indicators that show whether the design is working and how to respond when it is not.

TABLE XVII. OPERATIONAL INDICATORS FOR A DEPLOYMENT

Indicator	What it reveals	Typical response
95th-percentile and maximum hold	whether g and H suit current arrival patterns	tune g per channel or therapy class
Late-attachment share	documents arriving after release	raise g if attachments are frequent and costly
Identity review rate	ambiguity in incoming identity data	improve data quality at the source; recalibrate weights
Reported false merges	linking errors that reached a case	incident review; raise δ or τ_{hi}; audit weights
Insurance conflict rate by source	stale or inconsistent benefit data	correct at the source; verify first
Straight-through share by channel	where automation succeeds or fails	close the data gaps of the weakest channel
Technician queue wait	whether staffing matches released work	adjust staffing; avoid bursty release

F. Generalization

The same structure appears wherever related requests arrive separately but must be handled together: home-infusion orders and their supplies, equipment orders that accompany a prescription, laboratory orders that belong to one encounter, and prior authorization requests for the drugs of one regimen. The architecture carries over with a different milestone, different urgency rules and different fields in the identity score; the trade-offs quantified here, between hold and rework and between linking and safety, carry over with it.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

G. Limitations and threats to validity

Construct validity. The outcome measures are proxies. Labor is a linear function of cases, prescriptions and attachments, with assumed unit times, and ignores learning effects, interruptions and the extra effort that a large consolidated case may require. The straight-through predicate is a simplified version of real clearance rules, which also check formulary status, dose limits and prior history. Hold is measured as delay added before case creation, not as total turnaround, which depends mainly on work after intake. A conflict flag is counted as a benefit because it surfaces an inconsistency early, but resolving it costs technician time that the model does not count.

Internal validity. The grouping implementation was verified against a closed-form expression (Table VII), every policy was run on identical document streams with common random numbers, the first simulated day was discarded as warm-up, and each configuration was replicated 30 times. The identity scorer, its thresholds and the margin were set by hand; the threshold sweep in Fig. 9 shows how the identity results depend on them.

External validity. The arrival process, channel mix, attribute completeness, identity noise and unit times are assumptions chosen to be plausible, not measurements from any pharmacy, so absolute values should not be read as predictions. The conclusions that matter are comparative. Section VI-H shows that they are robust to the variations tested, but not to every possible variation; in particular, a setting in which most referrals carry a single prescription benefits less, as Section VI-E explains. The identity model uses synthetic names; real populations have other name distributions, and real scorers should be trained and calibrated on labeled data, with attention to accuracy across name origins.

Conclusion validity. With 30 paired replications, small differences reach statistical significance easily, so the tests in Table XVI establish only that the differences are systematic. Their size is better judged from the confidence intervals and the sensitivity analysis. Finally, the study does not model benefits verification and prior authorization after intake, where consolidation may bring further gains, such as one authorization request per episode, that are not counted here. A prospective evaluation on operational data, with appropriate privacy safeguards, is the natural next step.

VIII. CONCLUSION

Specialty referrals arrive as bursts of related prescriptions that per-prescription intake splits into separate cases. This paper formulated consolidation as keyed event correlation over a noisy stream and proposed an architecture that resolves patient identity through a federated entity service with an ambiguity margin, groups documents in session windows with a bounded hold and an urgency bypass, attaches late documents to open cases, and makes straight-through decisions on the consolidated case. In a simulation study, the design reduced technician minutes by 35.7% and raised the straight-through share from 21.1% to 35.0% relative to per-prescription intake, with a median hold of 20 minutes, a maximum of one hour, no hold for urgent therapies and no wrong links between twins. Future work includes learning windows per therapy class from arrival data, connecting the consolidated case to electronic prior authorization, and validating the results prospectively.

Conflict of interest. The author declares no competing interests. The study uses synthetic data only and does not describe the systems of any particular organization.

Data availability. All data in this study are synthetic; the simulation code is available from the author on reasonable request.

REFERENCES

[1] Pharmacy Quality Alliance, "Specialty Pharmacy Turnaround Time (SP-TAT): Measure rationale," PQA, 2021. [Online]. Available: https://www.pqaalliance.org/assets/docs/PQA_SP-TAT_Rationale.pdf

[2] M. H. Gabriel, C. M. Kotschevar, D. Tarver, V. Mastrangelo, L. Pezzullo, and P. J. Campbell, "Specialty pharmacy turnaround time impediments, facilitators, and good practices," *J. Manag. Care Spec. Pharm.*, vol. 28, no. 11, 2022, doi: 10.18553/jmcp.2022.28.11.1244.

[3] American Medical Association, "2022 AMA prior authorization (PA) physician survey," AMA, Chicago, IL, USA, 2023.

[4] American Society of Health-System Pharmacists, pharmacy administrator and pharmacy technician workforce surveys (Nov. 2021), reported in "Survey reveals severe shortage of pharmacy technicians in hospitals and health

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

systems," *News-Medical*, Mar. 15, 2022. [Online]. Available: https://www.news-medical.net/news/20220315/Survey-reveals-severe-shortage-of-pharmacy-technicians-in-hospitals-and-health-systems.aspx
[5] O. K. Odukoya, J. A. Stone, and M. A. Chui, "How do community pharmacies recover from e-prescription errors?," *Res. Social Adm. Pharm.*, vol. 10, no. 6, pp. 837–852, 2014, doi: 10.1016/j.sapharm.2013.11.009.
[6] Centers for Medicare & Medicaid Services, "Advancing interoperability and improving prior authorization processes (CMS-0057-F)," *Federal Register*, vol. 89, no. 27, p. 8758, Feb. 8, 2024.
[7] HL7 International, "FHIR Release 4 (v4.0.1)," 2019. [Online]. Available: https://hl7.org/fhir/R4/
[8] HL7 International, "Da Vinci Prior Authorization Support (PAS) FHIR Implementation Guide." [Online]. Available: https://hl7.org/fhir/us/davinci-pas/
[9] National Council for Prescription Drug Programs, *SCRIPT Standard Implementation Guide, Version 2017071*. Scottsdale, AZ, USA: NCPDP, 2017.
[10] T. Akidau, R. Bradshaw, C. Chambers, S. Chernyak, R. J. Fernández-Moctezuma, R. Lax, S. McVeety, D. Mills, F. Perry, E. Schmidt, and S. Whittle, "The Dataflow Model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing," *Proc. VLDB Endowment*, vol. 8, no. 12, pp. 1792–1803, 2015.
[11] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache Flink: Stream and batch processing in a single engine," *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.*, vol. 38, no. 4, pp. 28–38, 2015.
[12] D. Luckham, *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Boston, MA, USA: Addison-Wesley, 2002.
[13] G. Cugola and A. Margara, "Processing flows of information: From data stream to complex event processing," *ACM Comput. Surv.*, vol. 44, no. 3, Art. no. 15, 2012.
[14] G. Hohpe and B. Woolf, *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Boston, MA, USA: Addison-Wesley, 2003.
[15] J. Nagle, "Congestion control in IP/TCP internetworks," RFC 896, Internet Engineering Task Force, Jan. 1984.
[16] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *J. Amer. Statist. Assoc.*, vol. 64, no. 328, pp. 1183–1210, 1969.
[17] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *J. Amer. Statist. Assoc.*, vol. 84, no. 406, pp. 414–420, 1989.
[18] W. E. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage," in *Proc. Section on Survey Research Methods*, American Statistical Association, 1990, pp. 354–359.
[19] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Germany: Springer, 2012.
[20] O. Binette and R. C. Steorts, "(Almost) all of entity resolution," *Science Advances*, vol. 8, no. 12, Art. no. eabi8021, 2022.
[21] G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, "Blocking and filtering techniques for entity resolution: A survey," *ACM Comput. Surv.*, vol. 53, no. 2, Art. no. 31, 2020.
[22] Office of the National Coordinator for Health Information Technology, "Patient identification and matching: Final report," U.S. Dept. of Health and Human Services, Washington, DC, USA, Feb. 2014.
[23] B. H. Just, D. Marc, M. Munns, and R. Sandefer, "Why patient matching is a challenge: Research on master patient index (MPI) data discrepancies in key identifying fields," *Perspect. Health Inf. Manag.*, vol. 13, Spring, Art. no. 1e, 2016.
[24] S. J. Grannis, H. Xu, J. R. Vest, et al., "Evaluating the effect of data standardization and validation on patient matching accuracy," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 5, pp. 447–456, 2019, doi: 10.1093/jamia/ocy191.
[25] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching: A design space exploration," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2018, pp. 19–34.
[26] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, "Deep entity matching with pre-trained language models," *Proc. VLDB Endowment*, vol. 14, no. 1, pp. 50–60, 2020.
[27] The Pew Charitable Trusts, "Enhanced patient matching is critical to achieving full promise of digital health records," Oct. 2, 2018. [Online]. Available: https://www.pew.org/en/research-and-analysis/reports/2018/10/02/enhanced-patient-matching-critical-to-achieving-full-promise-of-digital-health-records
[28] M. Lenzerini, "Data integration: A theoretical perspective," in *Proc. 21st ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS)*, 2002, pp. 233–246.
[29] A. Halevy, A. Rajaraman, and J. Ordille, "Data integration: The teenage years," in *Proc. 32nd Int. Conf. Very Large Data Bases (VLDB)*, 2006, pp. 9–16.
[30] C. Richardson, *Microservices Patterns*. Shelter Island, NY, USA: Manning, 2018.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304474 |

[31] M. Kleppmann, *Designing Data-Intensive Applications*. Sebastopol, CA, USA: O'Reilly Media, 2017.
[32] P. Helland, "Idempotence is not a medical condition," *Commun. ACM*, vol. 55, no. 5, pp. 56–65, 2012.
[33] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A distributed messaging system for log processing," in *Proc. NetDB Workshop*, Athens, Greece, 2011.
[34] G. Varghese and T. Lauck, "Hashed and hierarchical timing wheels: Data structures for the efficient implementation of a timer facility," in *Proc. 11th ACM Symp. Operating Systems Principles (SOSP)*, 1987, pp. 25–38.
[35] H. Garcia-Molina and K. Salem, "Sagas," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 1987, pp. 249–259.
[36] Apache Software Foundation, "Kafka Streams DSL: Windowing," Apache Kafka documentation. [Online]. Available: https://kafka.apache.org/documentation/streams/developer-guide/dsl-api.html
[37] J. Dean and L. A. Barroso, "The tail at scale," *Commun. ACM*, vol. 56, no. 2, pp. 74–80, 2013.
[38] A. M. Law, *Simulation Modeling and Analysis*, 5th ed. New York, NY, USA: McGraw-Hill, 2015.
[39] J. D. C. Little, "A proof for the queuing formula: L = λW," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961.
[40] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.
[41] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Hoboken, NJ, USA: Wiley, 2003.
[42] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
[43] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
[44] U.S. Department of Health and Human Services, "Uses and disclosures of protected health information: General rules (minimum necessary)," 45 C.F.R. § 164.502(b).
[45] U.S. Department of Health and Human Services, "Security standards for the protection of electronic protected health information," 45 C.F.R. Part 164, Subpart C.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details