mathematics

MDPI

Article

Robust Bottleneck Prioritization from Event Logs Under Temporal Shift: A Wasserstein–CVaR Diagnostic Framework

Marian Pompiliu Cristescu

Faculty of Economic Sciences, Lucian Blaga University of Sibiu, Calea Dumbrăvii, no. 17, 550324 Sibiu, Romania; marian.cristescu@ulbsibiu.ro

Abstract

Bottleneck prioritization selects a small set of process transitions for investigation when historical delay patterns may change. This study examines frequency-weighted Wasserstein–Conditional Value-at-Risk scores for capped inter-event elapsed times. The decision snapshots contain only transitions whose origin and destination events precede the start of Q3. Under one-dimensional Wasserstein-1 ambiguity on a bounded interval, worst-case CVaR has an exact capped affine expression. Its saturation threshold is the transport budget needed to place the upper tail at the cap. A normalized radius-to-threshold ratio measures the remaining severity headroom, separately from portfolio discrimination. With a 90-day cap and a 0.95 tail level, 33 of 61 BPI Challenge 2019 candidates saturate. Original DRO and empirical tail select the same five transitions, capturing 66.97% of subsequent mean burden, compared with 70.23% for mean burden. The paired original-DRO-minus-empirical-tail difference is exactly zero. In BPI Challenge 2017, three of 41 candidates saturate. Clipping the radius below the threshold removes induced saturation, but the four severity-based methods retain identical top-five membership and capture 70.237% of mean burden. Calibration, observation-maturity and resampling sensitivities delimit these comparisons. The diagnostic identifies saturation of the risk component. It does not establish that saturation necessarily changes the selected portfolio or that clipping improves future prioritization.

Keywords: process mining; event logs; bottleneck prioritization; temporal distribution shift; distributionally robust optimization; Wasserstein distance; Conditional Value-at-Risk; CVaR saturation; robust decision support

MSC: 90C17; 90C15; 90C90; 68U35

Check for updates

Academic Editor: Iraklis Varlamis

Received: 18 September 2026
Revised: 8 October 2026
Accepted: 26 October 2026
Published: 29 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Process mining uses event data recorded by enterprise systems to reconstruct, monitor, and improve operational processes [1]. An event log can be understood as a digital history of process execution: each event records that an activity occurred for a particular case at a particular time, and the ordered events of one case form a trace [1]. A directly follow transition is the pair formed when one recorded activity is immediately followed by another within the same trace. This local transition view is useful for performance analysis because large elapsed times between consecutive activities can reveal hand-offs or segments that deserve investigation, although the elapsed interval should not automatically be interpreted as pure waiting or processing time [1,2].

From a managerial perspective, the central problem is prioritization. A process may contain hundreds of distinct transitions, while analysts and process owners can investigate

or redesign only a small subset at a time. Bottleneck analysis therefore seeks to identify process segments that contribute disproportionately to delay, congestion, or poor cycle-time performance [2,3]. A useful ranking must consider both severity and prevalence: a rare transition with a very long delay may matter less to total process burden than a moderately slow transition that occurs thousands of times. This motivates transition-level burden scores that combine frequency with a delay statistic rather than ranking transitions by frequency or delay alone.

Historical rankings can become unreliable when workloads, business rules or process behavior change. Such changes are studied as process or concept drift [4–6]. Here, temporal shift concerns the distribution of elapsed time for a recorded transition. The comparison uses a ranking constructed from a 1 July decision snapshot and evaluates it on subsequent Q3 transition origins. The snapshot restricts both event timestamps, not only the timestamp of the origin [4–7].

Conditional Value-at-Risk summarizes the mean severity of the upper probability tail [8]. At a level of 0.95, the empirical calculation integrates the largest 5% of probability mass, including a fractional observation when necessary. Distributionally robust optimization evaluates losses over an ambiguity set rather than one reference distribution [9,10]. The Wasserstein-1 distance measures the minimum probability-mass-times-distance transport cost between distributions. Related Earth Mover measures compare distributions in stochastic process mining [11,12]. In this model, a larger radius admits a wider range of capped-delay distributions.

The bounded outcome imposes a limit on robust tail severity. Worst-case CVaR cannot exceed the cap C. Once the ambiguity budget can transport the required upper-tail mass to C, that risk value stops increasing. Different transitions can then share the same robust severity, although their frequency-weighted scores may remain different. Whether saturation changes a top-K selection is therefore a separate empirical question.

BPI Challenge 2019 [13] provides the development setting and BPI Challenge 2017 [14] the cross-log evaluation. The analysis combines historical block calibration, availability-aware decision snapshots and a transport-based saturation diagnostic. The diagnostic is assessed against both risk values and selected portfolios. The two logs are evaluated retrospectively under a common specification. Their Q3 outcomes do not enter candidate eligibility, radius calculation or historical scores. The cross-log comparison is not treated as evidence of universal ranking superiority.

1.1. Related Work, Conceptual Positioning, and Research Gap

Process-performance analysis identifies slow or high-impact segments and investigates their causes. Lashkevich et al. [2] decompose waiting time at activity transitions, while Tang et al. [3] combine clustering with knowledge graphs for bottleneck analysis. The present outcome remains inter-event elapsed time. The objective is to select transitions for investigation under temporal uncertainty, without inferring a causal waiting mechanism or an intervention effect.

Prescriptive process monitoring addresses a related but different decision problem. The field has moved from outcome prediction toward recommendations about interventions, treatment assignment, causal effects, resource constraints, and learned decision policies [15–19]. These methods ask what action should be taken for an ongoing case, often with intervention costs or treatment effects. Our problem is one level earlier: before a process owner can design or evaluate an intervention, a limited set of transitions must often be selected for investigation. We therefore treat high-ranked transitions as candidates for process-improvement attention, not as proof that intervening on those transitions will causally improve the process.

Concept-drift research examines changes in process behavior [4–7]. Stochastic process mining uses transport distances to compare distributions [11,12], and Wasserstein DRO constructs ambiguity sets around empirical observations [9,10]. CVaR supplies the tail-risk functional [8]. Worst-case risk measures for transformed or limited losses are studied by Cai et al. [20]. Here, the ambiguity set is placed directly on the capped-delay distribution, rather than on an uncapped loss before transformation. The application-specific contribution is the joint assessment of saturation, transition frequency and selected bottleneck portfolios across decision snapshots. The mathematical characterization follows the bounded one-dimensional risk structure.

1.2. Research Questions and Contribution Logic

RQ1. How does a transition-specific Wasserstein–CVaR ranking compare with simpler frequency, mean-burden, and empirical-tail rankings under an out-of-time temporal shift?

RQ2. Under bounded inter-event elapsed time, when does the calibrated Wasserstein radius exhaust severity headroom, and how does that saturation relate to weighted-score discrimination and portfolio membership?

RQ3. How does the saturation diagnostic behave in a different event log, and does radius clipping alter the resulting out-of-time portfolios?

The evaluation distinguishes future burden capture from the diagnostic of bounded risk. A correct saturation threshold can coexist with unchanged portfolio membership or with no improvement in holdout performance. These outcomes answer different questions and are reported separately.

2. Materials and Methods

2.1. Event Logs and Study Unit

The development log is BPI Challenge 2019, a public purchase-to-pay event log from a large Netherlands-based multinational [13]. In process-mining terminology, a case is one process instance, an event is one recorded activity occurrence with a timestamp, and the chronologically ordered events of a case form its trace [1]. Our analytical unit is a directly follow transition: two activities that appear consecutively in the same trace [1]. For every such pair we measure inter-event elapsed time as destination timestamp minus origin timestamp. This interval is deliberately interpreted as an observational process-performance signal. In BPI 2019, the absence of paired lifecycle start/complete timestamps means that it cannot be decomposed into pure service time, processing time, or queue waiting time [2].

The external validation log is BPI Challenge 2017, which records a Dutch financial institution's loan-application process [14]. Because this log contains lifecycle information, the primary external analysis uses lifecycle COMPLETE events only and forms directly follow transitions between consecutive COMPLETE events. This harmonizes the unit of analysis across the two logs: in both cases the ranking concerns the elapsed time between consecutive recorded process events rather than a causal decomposition of activity duration. Table 1 summarizes the source logs, decision cutoffs, and outcome windows.

Table 1. Event-log sources, decision cutoffs and outcome observation windows.

Design Feature	BPI Challenge 2019	BPI Challenge 2017
Process domain	Purchase-to-pay	Loan applications
Raw size	251,734 cases, 1,595,923 events	31,509 cases, 1,202,267 events
Analytical event population	251,268 cases, 1,587,466 valid events	475,306 lifecycle COMPLETE events
Primary event semantics	Consecutive recorded events	Consecutive COMPLETE events
Historical origin window	January to June 2018	January to June 2016

Table 1. *Cont.*

Design Feature	BPI Challenge 2019	BPI Challenge 2017
Decision cutoff, UTC	1 July 2018, 00:00	1 July 2016, 00:00
Outcome origin window	July to September 2018	July to September 2016
Observation cutoff, UTC	18 January 2019, 13:34	1 February 2017, 14:00:30.347
Primary cap	2160 h, 90 days	2160 h, 90 days
CVaR level	0.95	0.95
Primary representation	Original empirical support	Original empirical support
Available historical pairs	601,083	191,097
Candidate transitions	61	41

2.2. BPI 2019 Cohort Construction and Observation-Edge Control

The BPI 2019 analytical cohort contains cases whose earliest recorded event is in 2018. Events with timestamps outside 2017–2019 are excluded individually. A later anomalous event does not exclude the earlier history of its case. This rule retains 251,268 cases and 1,587,466 valid events. For each case, events are sorted by timestamp, with original file order resolving ties. Zero elapsed times are retained. The supplied log contains 231,420 timestamp-tied pairs in this cohort, so ordering within a tie remains a recording convention rather than an identifiable temporal ordering.

The primary BPI 2019 observation cutoff is 18 January 2019 at 13:34 UTC. Sparse later timestamps occur in the supplied record, but do not extend this primary cutoff. BPI 2017 is observed through 1 February 2017 at 14:00:30.347 UTC. The primary cap is 90 days. Q3 origins have at least 90 days of calendar follow-up before these cutoffs. The 90-day and 120-day comparison uses an identical origin window with at least 120 days of follow-up. In BPI 2019, this window ends before 20 September 2018 at 13:34 UTC. In BPI 2017, it includes the full Q3 origin window. Sufficient calendar age does not establish that every successor has been recorded.

2.3. Candidate Transitions and Decision Snapshots

Historical origins lie in January–June 2018 for BPI 2019 and January–June 2016 for BPI 2017. Both the origin and destination must precede 1 July at 00:00 UTC of the respective year. No completed duration or destination label recorded after that cutoff enters the historical model. A candidate requires at least 500 available historical pairs and valid blocks in at least 60% of the calendar blocks, rounded upward. A valid block contains at least 20 pairs. The primary 14-day partition contains 13 blocks and therefore requires eight valid blocks. The same coverage fraction is used for alternative block lengths and maturity windows. This yields 61 BPI 2019 and 41 BPI 2017 candidates.

The historical snapshots contain 601,083 available BPI 2019 pairs and 191,097 BPI 2017 pairs. At the respective decision cutoffs, 58,811 and 2097 historical origin events have successors recorded only later. They are unavailable to the ranking model. Origins with no successor anywhere in the supplied record are reported separately. They may be terminal events or incomplete observations, and are not assigned a destination transition. Historical maturity checks restrict origins to dates at least 90 or 120 days before the same decision cutoff. These checks assess completion-related selection, but do not identify unobserved successor labels. The estimand is burden among recorded-successor transitions, conditional on the stated observation rule. Availability is inferred from event timestamps because ingestion timestamps are not supplied. Figure 1 summarizes the decision cutoffs, fixed portfolios, and subsequent Q3 evaluation.

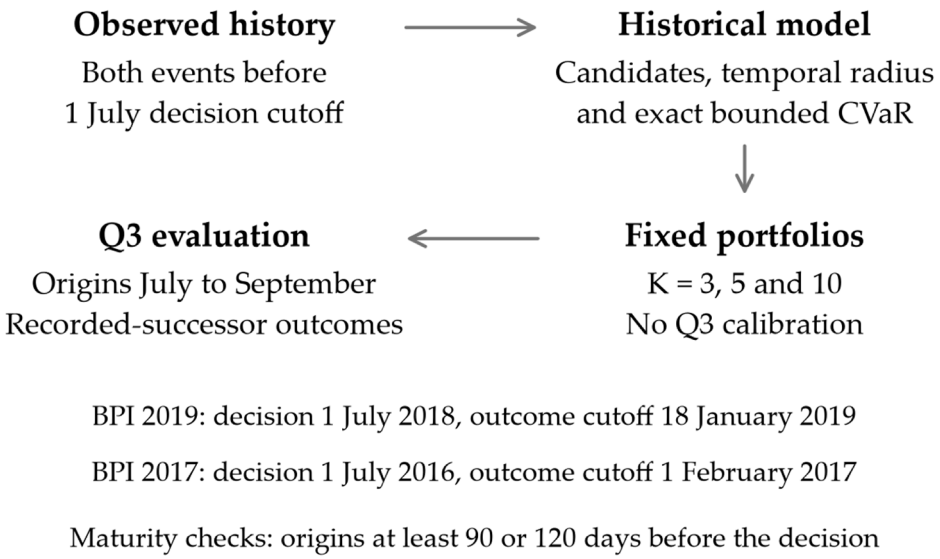


Figure 1. Decision snapshots and recorded-successor evaluation in the two event logs.

2.4. Capped-Delay Distributions and Ranking Scores

For transition j, Δ_{ji} is the non-negative difference between the destination and origin timestamps. The capped outcome uses $C = 2160$ h, with 120 days assessed separately. An observed duration above C contributes C to the risk calculation. This bounds the modeled outcome but does not replace a missing successor or a censored duration.

$$Y_{ji} = \min(\Delta_{ji}, C), 0 \leq Y_{ji} \leq C \quad (1)$$

Let n_j denote the number of available historical pairs of transition j. Its historical frequency is $f_j = \frac{n_j}{\sum_k n_k}$, where the denominator includes all available historical transitions, not only eligible candidates. Mean burden multiplies frequency by mean capped delay. Empirical tail uses empirical CVaR. Original DRO uses worst-case CVaR over the temporal ambiguity set. The frequency component is held at its historical estimate in all three severity-based scores. Table 2 defines the historical transition scores.

Table 2. Historical transition scores used for fixed-portfolio evaluation.

Method	Transition Score
Frequency	$S_j = f_j$
Mean burden	$S_j = f_j \cdot E_{\hat{P}_j}[Y]$
Empirical tail	$S_j = f_j \cdot \mathrm{CVaR}_\alpha\left(\hat{P}_j\right)$
Original Wasserstein-DRO	$S_j = f_j \cdot R_j(\varepsilon_j)$, where Rj is worst-case CVaR

The portfolio sizes are $K \in \{3, 5, 10\}$, with K = 5 the primary size. Ties are resolved by transition name in ascending order. A portfolio is a membership set, so permutations of its internal ranking do not change captured burden.

2.5. Historical Wasserstein Calibration

The capped historical sample defines $\hat{P}_j$. For each valid calendar block, the Wasserstein-1 distance compares the block distribution with the pooled historical distribution [9,10]. The primary partition uses consecutive 14-day blocks anchored at 1 January. All available historical pairs contribute to the pooled reference, including pairs in blocks below the

minimum block size. The radius is the empirical 90th percentile of valid-block distances, using linear interpolation between order statistics.

$$\varepsilon_j = Q_{0.90}\left(\left\{W_1\left(\hat{P}_{j,b},\ \hat{P}_j\right) :\ \text{b is a valid H1 block}\right\}\right) \tag{2}$$

The radius is a descriptive scale of within-history variation. It mixes temporal change with finite-sample variation and is not a 90% confidence or coverage guarantee for the next quarter. Calibration checks use 7-day and 28-day blocks, block-size minima of 50 and 100, and a leave-one-block-out reference. These checks retain the primary candidates wherever at least one qualifying block permits radius estimation and report strict eligibility separately. A candidate with no qualifying block is unavailable in that specification. A finite-sample benchmark permutes capped delays among the original block slots 100 times for each candidate. It preserves the pooled distribution and block sizes, but assumes exchangeable observations and does not preserve case dependence.

2.6. Wasserstein-Robust CVaR

For a probability distribution P supported on $[0, C]$, CVaR at level α has the Rockafellar–Uryasev variational representation [8]. For the elapsed-time loss and $\alpha = 0.95$, it averages the upper 5% of probability mass, with fractional atom weighting when required.

$$\mathrm{CVaR}_\alpha(P) = \min_{t\in[0,C]}\left\{t + \frac{1}{1-\alpha}\,E_P\left[(Y-t)_+\right]\right\} \tag{3}$$

The ambiguity set contains distributions within Wasserstein-1 distance ε_j of the historical distribution $\hat{P}_j$, with support restricted to $[0, C]$ [9,10].

$$\mathcal{D}_j\left(\varepsilon_j\right) = \left\{Q :\ W_1\left(Q,\ \hat{P}_j\right) \le \varepsilon_j,\ supp(Q) \subseteq [0, C]\right\} \tag{4}$$

The robust functional evaluates the largest CVaR within the bounded Wasserstein ambiguity set [9,10]. The variational expression uses $\alpha = 0.95$ and thresholds in $[0, C]$. The compactness of the support and convexity of the ambiguity set permit the minimax interchange described in Appendix A.

$$R_j\left(\varepsilon_j\right) = \min_{t\in[0,C]}\left\{t + \frac{1}{1-\alpha}\sup_{Q\in\mathcal{D}_j(\varepsilon_j)} E_Q\left[(Y-t)_+\right]\right\} \tag{5}$$

For the stated one-dimensional model with absolute-distance transport cost, the exact calculation on the original empirical support is

$$R_j\left(\varepsilon_j\right) = \min\left\{C,\ \mathrm{CVaR}_\alpha\left(\hat{P}_j\right) + \frac{\varepsilon_j}{1-\alpha}\right\}$$

The radius acts on the capped-delay distribution itself. The formula does not apply unchanged to a higher-order Wasserstein distance, an ambiguity set with additional moment restrictions, or ambiguity defined on uncapped delays before transformation. Original-support empirical CVaR is used by both empirical tail and original DRO. Linear programs independently check the formula rather than determine the primary score.

2.7. Bounded-Support Saturation Threshold

Bounded support gives $0 \le Y \le C$ and prevents robust CVaR from exceeding C. Let $m_C = \hat{P}(Y = C)$ be the empirical mass at the cap. If this mass is at least $1 - \alpha$, empirical CVaR is already saturated. Otherwise, the additional probability mass needed at C is

$$\delta = \max\{0, 1 - \alpha - m_C\} \tag{6}$$

The saturation threshold is the minimum transport cost required to place that missing mass at C. For support points below the cap, moving mass from the largest point is cheapest. Writing the support as $x_1 < \cdots < x_m = C$ and appending C with zero mass if necessary gives

$$\begin{aligned} \varepsilon_{sat} &= \min_{\{r_i\}_{i<m}} \textstyle\sum_{i<m} r_i (C - x_i) \\ &\text{s.t. } 0 \le r_i \le p_i,\ i < m \\ &\textstyle\sum_{i<m} r_i = \delta \end{aligned} \tag{7}$$

The optimal allocation moves probability from the upper empirical tail toward C, splitting an atom when only part of its mass is required. For the generalized inverse distribution function, the equivalent expression is

$$\varepsilon_{sat}\left(\hat{P},\ \alpha,\ C\right) = \begin{cases} 0 & m_C \ge 1 - \alpha \\ \int_{\alpha}^{1-m_C} \left[C - \hat{F}^{-1}(u)\right] du & m_C < 1 - \alpha \end{cases} \tag{8}$$

Proposition. For $0 < \alpha < 1$ and the ambiguity set in (4), $R(\varepsilon) = C \Leftrightarrow \varepsilon \ge \varepsilon_{\text{sat}}$. The transport expression in (7), the quantile expression in (8) and the CVaR-headroom expression below coincide. Appendix A gives the proof and treats empirical atoms and the zero-threshold case [8,20].

$$\varepsilon_{\text{sat},j} = (1 - \alpha)\left(C - \text{CVaR}_{\alpha}\left(\hat{P}_j\right)\right)$$

For positive thresholds, $\rho_j = \frac{\varepsilon_j}{\varepsilon_{\text{sat},j}}$. A ratio below one leaves severity headroom, while $\rho_j \ge 1$ predicts a saturated risk component. When the threshold is zero, the ratio is undefined and empirical saturation is reported separately. Even if several robust risks equal C, their weighted scores remain different whenever their historical frequencies differ. Portfolio discrimination must therefore be examined directly.

2.8. Holdout Evaluation, Bootstrap, and Sensitivity Analysis

Captured mean burden is the primary operational outcome. Weighted transition-tail burden is secondary. The Q3 outcome population contains origins from 1 July to 1 October, excluding the latter endpoint, with successors recorded by the observation cutoff. For a selected set A, captured mean burden is the sum of its Q3 capped durations divided by the sum over all observed Q3 pairs. Captured tail burden uses the sum of transition-specific Q3 frequencies multiplied by their empirical CVaRs. Its denominator includes all observed Q3 transition types. It is a weighted sum of transition CVaRs, not the CVaR of the mixture of all Q3 delays.

$$S_{\text{Mean}}(A) = \frac{\sum_{j \in A} n_j^{\text{Q3}} \hat{\mu}_j}{\sum_{l \in U} n_l^{\text{Q3}} \hat{\mu}_l}$$

$$S_{\text{Tail}}(A) = \frac{\sum_{j \in A} n_j^{\text{Q3}} \text{CVaR}_{\alpha}\left(\hat{P}_j^{\text{Q3}}\right)}{\sum_{l \in U} n_l^{\text{Q3}} \text{CVaR}_{\alpha}\left(\hat{P}_l^{\text{Q3}}\right)}$$

Here, U is the set of transition types with recorded Q3 pairs. The transition count is n_j^{Q3}, its empirical capped-delay distribution is $\hat{P}_j^{\text{Q3}}$ and its mean is $\hat{\mu}_j$. Reported shares are 100 times these ratios. A historical candidate absent from Q3 contributes zero burden, without assigning it an estimated Q3 CVaR. A transition that is new in Q3 remains in the denominator but is not retrospectively admitted to a historical portfolio. The eligible oracle selects the best K transitions from the same historical candidate set. The global oracle selects from all Q3 types and is reported separately. Regret is the oracle share minus the method share in percentage points. Kendall tau-b compares historical scores with Q3 mean or tail burden for historically eligible transitions observed in Q3. Portfolio overlap is Jaccard intersection over union.

Q3 uncertainty uses 1000 common resamples for every method. The primary partition has thirteen full seven-day origin blocks and one final one-day block. Each replicate draws fourteen blocks with replacement, carries all their observed pairs, and recomputes transition-specific Q3 CVaR. Case observations can occur in more than one calendar block. Sensitivities use fourteen-day and twenty-eight-day partitions, and case clusters that assign the same resampling weight to all included Q3 pairs of a case. Percentile intervals use the 2.5th and 97.5th percentiles. Identical membership sets are summed in canonical transition order and have exactly zero paired differences.

Historical selection stability uses 500 resamples of the thirteen fourteen-day blocks. Candidate eligibility, historical frequency, empirical mean, CVaR, pooled reference, temporal radius and saturation-aware radius are refitted in each replicate. A repeated historical block retains its observations and counts as a repeated block occurrence. The stability outputs report eligibility probability, top-K inclusion probability and Jaccard overlap with the original snapshot selection. These refits concern selection from the observed historical population. Q3 confidence intervals remain conditional on the fixed primary portfolios.

2.9. Cross-Log Saturation-Aware Evaluation

BPI 2017 supplies a different process domain under the same decision-snapshot, candidate and evaluation rules. Both logs are analyzed retrospectively under the same specification. The parameters are common across logs, and the Q3 outcomes do not enter the historical calculations. One-factor sensitivities describe dependence on the specification without selecting a preferred value from Q3 performance.

Saturation-aware clipping limits the temporal radius to a fraction of the positive saturation threshold. The operational purpose is to retain severity headroom when the capped score is used to prioritize investigation. The smaller ambiguity set excludes some distributions admitted by the temporal radius. It therefore exchanges part of the original protection for score resolution, rather than preserving both unchanged.

$$\varepsilon_{SA} = \begin{cases} \min\left(\varepsilon_{temporal},\ \kappa\varepsilon_{sat}\right) & \varepsilon_{sat} > 0 \\ 0 & \varepsilon_{sat} = 0 \end{cases}, \ \kappa = 0.90 \tag{9}$$

The primary value $\kappa = 0.90$ leaves 10% of the CVaR-to-cap headroom whenever clipping is active. It is a transparent design margin, not an estimated optimal parameter or a probabilistic guarantee. Isolated sensitivity values are 0.50, 0.75, 0.95 and 0.99. If empirical CVaR is already C, the threshold is zero and a zero radius leaves the risk at C. Clipping can remove induced saturation but cannot undo empirical saturation.

2.10. Computational Checks and Reproducibility

The original empirical support is the primary representation. Grid sensitivities round capped delays to the nearest 12 h or 24 h with ties to the even grid index. Zero and C are retained exactly. The full empirical distribution and the block distributions are mapped to the same grid in each sensitivity, so the corresponding tail and DRO scores use identical representations. Grid sensitivity reports CVaR, robust risk, saturation threshold, finite threshold ratio and classification, rather than only top-K membership.

The exact formula is checked against a dual LP for every primary candidate on a common 24 h support at zero radius, half the saturation threshold, the threshold, twice the threshold and the temporal radius, with duplicate budgets omitted. The threshold is also evaluated independently by greedy transport on the original support. Synthetic distributions include a fractional upper-tail atom and empirical mass already sufficient to saturate. The manifest supplies the full code hashes for version 1.1.0, as detailed in the Supplementary Materials. Table A9 maps the manuscript results to their aggregate outputs.

3. Results

3.1. BPI 2019 Process Structure and Candidate Set

The analytical BPI 2019 cohort has 42 activities. The historical decision snapshot contains 601,083 observed pairs, of which 61 transition types meet the primary candidate rule. The Q3 population contains 352,493 pairs across 378 transition types. The candidate set accounts for 96.23% of Q3 mean burden and 96.34% of weighted tail burden. These coverage values are distinct from the share captured by a five-transition portfolio.

3.2. BPI 2019 Out-of-Time Ranking Performance

Empirical tail and original DRO select the same five BPI 2019 transitions: Record Invoice Receipt to Clear Invoice, Create Purchase Order Item to Vendor Creates Invoice, Record Goods Receipt to Record Invoice Receipt, Vendor Creates Invoice to Record Invoice Receipt, and Create Purchase Order Item to Record Goods Receipt. Their internal orders differ, as shown in Table A7. Both capture 66.97% of Q3 mean burden and 60.94% of weighted tail burden. Mean burden captures 70.23% and 59.23%, respectively. Frequency captures 55.55% and 46.36%. Thus, the mean-oriented rule has the larger mean-burden share, while the tail-oriented rules have the larger tail-burden point estimate. Table 3 reports the primary BPI 2019 K = 5 results. Figure 2 compares their mean and weighted tail-burden capture.

Table 3. BPI 2019 primary K = 5 results. Shares use all recorded Q3 transitions, and regret uses the eligible-candidate oracle. Oracle mean and tail columns use separately optimized portfolios.

Method	Mean Share (%)	Tail Share (%)	Mean Regret (pp)	Tail Regret (pp)
Frequency	55.547	46.363	16.426	14.985
Mean burden	70.229	59.234	1.745	2.114
Empirical tail	66.971	60.942	5.002	0.407
Original DRO	66.971	60.942	5.002	0.407
Eligible oracle	71.973	61.348	0	0
Global oracle	71.973	61.348	Not applicable	Not applicable

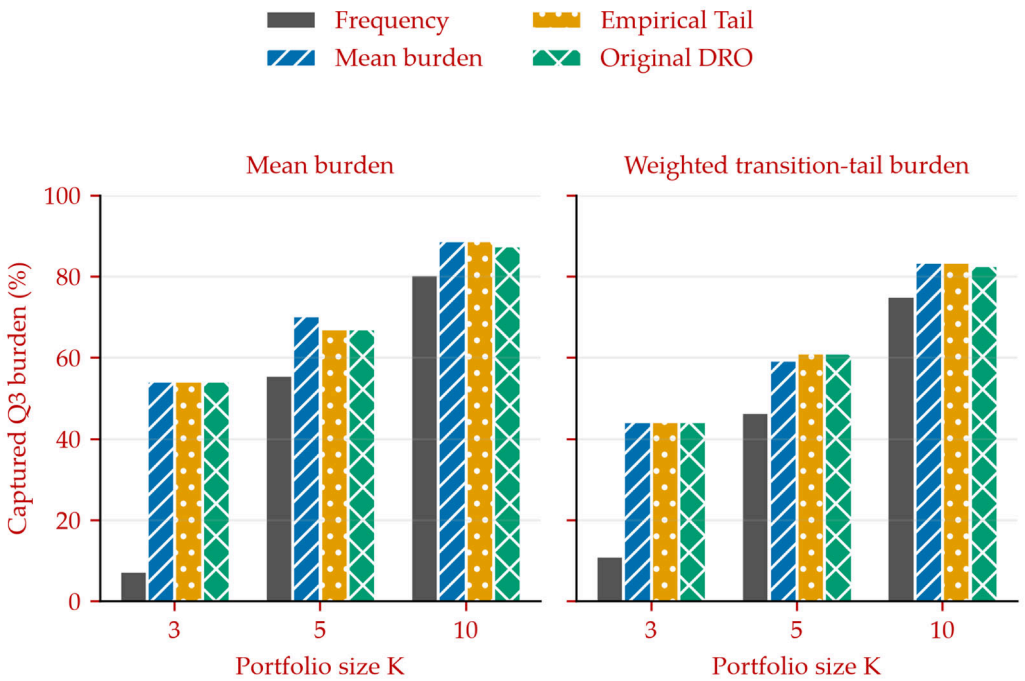


Figure 2. Mean and weighted transition-tail burden capture differ because the historical scores target different aspects of delay.

The observed original-DRO-minus-mean-burden difference is −3.26 percentage points, with a paired weekly block 95% percentile interval of [−5.99, −0.16]. The corresponding tail-burden difference is 1.71 points, with an interval of [−0.42, 3.94]. Original DRO exceeds frequency in mean burden by 11.42 points, with an interval of [9.94, 13.30]. Original DRO minus empirical tail is exactly zero in every common resample for both outcomes, with an interval of [0, 0]. Table A5 gives alternative dependence schemes.

3.3. Saturation and Portfolio Membership

On the available historical sample, three of 61 candidates are empirically saturated and 30 additional candidates saturate under the temporal radius. The total is 33 of 61, or 54.1%. The median finite radius-to-threshold ratio is 1.132. All five original-DRO selections have risk equal to 2160 h, so their robust scores are proportional to historical frequency. The score of the fifth selected transition is 134.13 h and the largest unselected score is 77.85 h. This gap shows why saturated severity values need not imply an ambiguous top-five selection. Figure 3 depicts how robust risk reaches the cap as the normalized radius reaches its threshold.

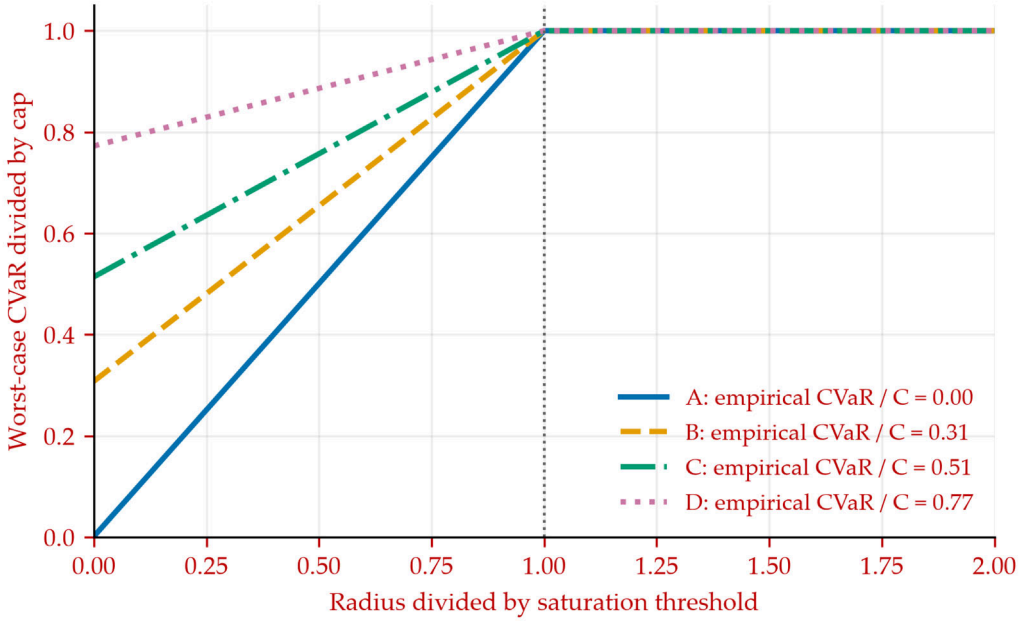


Figure 3. Positive-threshold candidates reach the bounded risk endpoint at a normalized radius of one. Empirically saturated candidates have no finite threshold ratio.

The exact risk formula, greedy transport threshold and numerical dual linear programs agree within numerical tolerance. Zero radius reproduces empirical CVaR on the same representation, and a positive threshold places the robust risk at C. The numerical agreement verifies the implementation. It is not independent evidence that the diagnostic improves a portfolio.

Across radius multipliers 0, 0.5, 1, 1.5 and 2, the BPI 2019 saturation count ranges from 3 to 42, while top-three and top-five membership remain unchanged. A piecewise-linear boundary audit confirms the same membership over the interval between these settings. At K = 10, the zero-radius portfolio includes Create Purchase Order Item to Change Quantity, whereas the primary DRO portfolio includes Record Goods Receipt to Remove Payment Block. The former portfolio captures 88.69% of Q3 mean burden and the latter 87.43%. This 1.26-point difference is a selection effect at the larger portfolio size. It does not explain the top-five difference from mean burden, because the zero-radius tail rule already selects the same top five.

3.4. Cross-Log Saturation Regimes

The available BPI 2017 historical sample contains 191,097 pairs. Forty-one candidates meet the primary rule. None is empirically saturated and three become saturated under the original radius. The median finite ratio is 0.00615. With the saturation-aware radius, no candidate is saturated. This is a low-saturation regime relative to BPI 2019 under the same cap and tail level, not evidence that the risk calculation is independent of the specification. Figure 4 places both logs in their radius-to-threshold regimes. Table 4 reports the corresponding saturation counts.

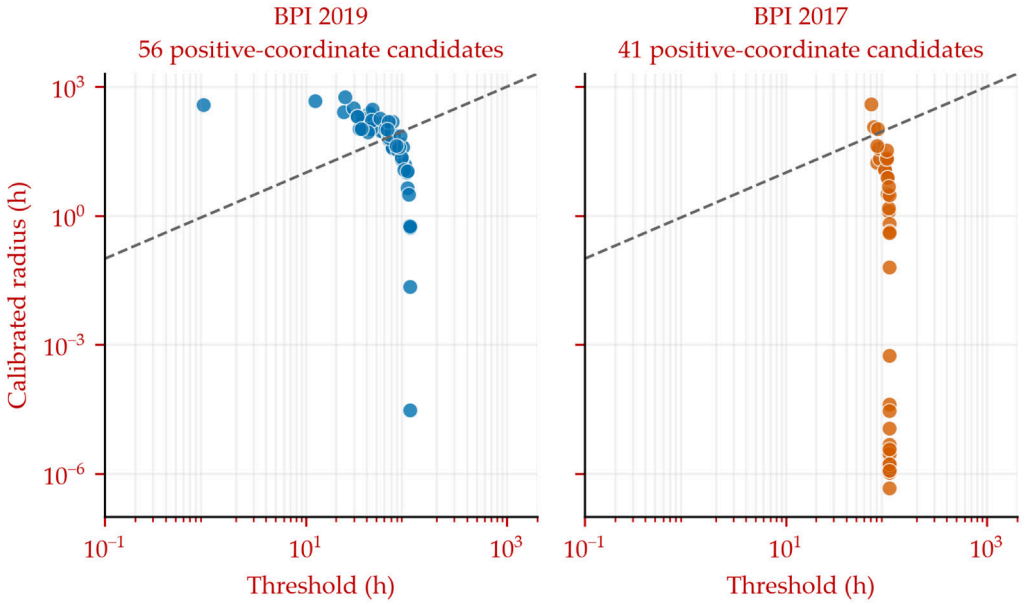


Figure 4. The two logs occupy different radius-to-threshold regimes. The diagonal marks equal transport radius and saturation threshold. Zero coordinates are omitted from the logarithmic axes.

Table 4. Saturation of the risk component on the available historical samples.

Diagnostic	BPI 2019	BPI 2017
Candidates	61	41
Empirically saturated	3	0
Induced by original DRO	30	3
Original-DRO total	33 (54.1%)	3 (7.3%)
SA-DRO empirical/induced	3/0	0/0
Median finite ratio	1.132	0.00615

3.5. BPI 2017 Out-of-Time Performance

In BPI 2017, mean burden, empirical tail, original DRO and saturation-aware DRO select the same five-transition membership set. The transitions are A_Complete to A_Cancelled, A_Complete to A_Validating, A_Concept to A_Accepted, A_Complete to O_Create Offer, and A_Incomplete to A_Validating. Their within-set ranks are listed in Table A8. Each method captures 70.237% of Q3 mean burden and 60.461% of weighted tail burden. Frequency captures 7.327% and 17.964%. The eligible and global mean oracles both capture 70.649%, while the corresponding tail oracles capture 60.461%. Table 5 reports the BPI 2017 primary K = 5 shares and oracle regrets.

Table 5. BPI 2017 primary K = 5 results, with mean and tail regrets against the eligible-candidate oracles. Oracle mean and tail columns use separately optimized portfolios.

Method	Mean Share (%)	Tail Share (%)	Mean Regret (pp)	Tail Regret (pp)
Frequency	7.327	17.964	63.322	42.496
Mean burden	70.237	60.461	0.412	0.000
Empirical tail	70.237	60.461	0.412	0.000
Original DRO	70.237	60.461	0.412	0.000
Saturation-aware DRO	70.237	60.461	0.412	0.000
Eligible oracle	70.649	60.461	0	0
Global oracle	70.649	60.461	Not applicable	Not applicable

Saturation-aware DRO exceeds frequency by 62.91 mean-burden percentage points, with a paired weekly block interval of [61.38, 64.44]. Its differences from mean burden, empirical tail and original DRO are exactly zero for both outcomes under every common resample. Removing three induced saturations does not change the primary top-five portfolio or establish incremental decision value.

The cross-log comparison supports the distinction between a saturated risk component and a useful selected set. The threshold identifies the former exactly. Out-of-time portfolio comparisons assess the latter and do not show an advantage for clipping at K = 5. Figure 5 shows the coincident mean-burden capture for the BPI 2017 severity-based methods.

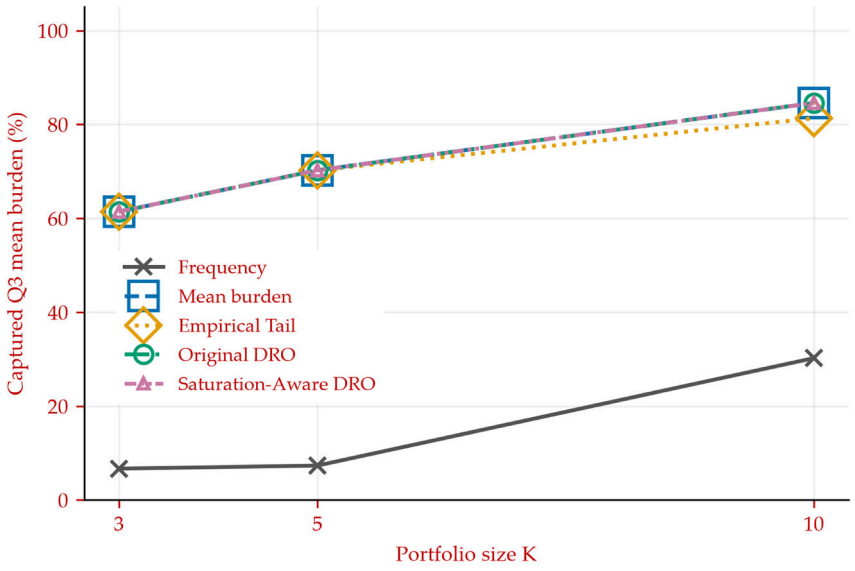


Figure 5. Severity-based BPI 2017 methods share the same K = 5 outcome. Distinct line patterns and nested hollow markers retain the coincident coordinates.

3.6. Calibration and Specification Sensitivity

Alternative calibration partitions change the saturation diagnostic more than the primary top-five membership. In BPI 2019, seven-day blocks produce 36 saturated candidates and twenty-eight-day blocks produce 35, compared with 33 under fourteen-day blocks. The corresponding BPI 2017 totals are 2, 3 and 3. The excluded-block reference gives 36 and 3. All these original-DRO K = 5 sets coincide with their primary counterparts. Raising the block minimum to 100 leaves no qualifying block for 11 BPI 2019 and four BPI 2017 candidates. Those candidates are unavailable in that sensitivity, rather than assigned an artificial radius. Table 6 reports the one-factor calibration sensitivities.

Table 6. One-factor calibration sensitivities on estimable primary candidates. J5 is Jaccard overlap with the primary original-DRO top five.

Log	Calibration	Estimable N	Saturated N	Median Ratio	J5
2019	14 days, minimum 20	61	33	1.13227	1.000
2019	7 days, minimum 20	61	36	1.74105	1.000
2019	28 days, minimum 20	61	35	1.15810	1.000
2019	14 days, minimum 50	61	32	1.06777	1.000
2019	14 days, minimum 100	50	27	1.01563	1.000
2019	14 days, excluded-block reference	61	36	1.20141	1.000
2017	14 days, minimum 20	41	3	0.00615	1.000
2017	7 days, minimum 20	41	2	0.00786	1.000
2017	28 days, minimum 20	41	3	0.00519	1.000
2017	14 days, minimum 50	41	2	0.00615	1.000
2017	14 days, minimum 100	37	2	0.00615	1.000
2017	14 days, excluded-block reference	41	3	0.00658	1.000

The permutation benchmark places the observed radius above its within-null 95th percentile for 51 of 61 BPI 2019 candidates and 24 of 41 BPI 2017 candidates. The median finite observed-to-null-median ratio is 3.10 and 1.51, respectively. Temporal variation is therefore not interchangeable with this exchangeable-observation benchmark. The comparison does not identify how much variation is due to case dependence, and it supplies no future-coverage guarantee.

In BPI 2019, increasing the cap from 90 to 120 days reduces the saturation total from 33 to 30 and the median finite ratio from 1.132 to 0.878. Both top-five sets are unchanged. Their shares are compared on the common early-Q3 origin window, not against the full-Q3 primary share. In BPI 2017, saturation falls from 3 to 2 and the ratio from 0.00615 to 0.00460. Table A3 isolates kappa. Across 0.50, 0.75, 0.90, 0.95 and 0.99, the external clipped top five and its 70.237% mean share remain unchanged. Clipping is active for four candidates at 0.50 and three at the other settings. Table 7 reports cap-sensitivity results on the shared mature-origin window.

Table 7. Cap sensitivity with the same sufficiently aged Q3 origin window within each log. The historical sample, alpha and other parameters are held fixed.

Log	Cap (Days)	Saturated N	Median Ratio	DRO Mean Share (%)	J5
2019	90	33	1.13227	67.918	1.000
2019	120	30	0.87844	67.529	1.000
2017	90	3	0.00615	70.237	1.000
2017	120	2	0.00460	70.207	1.000

Grid invariance of a selected set does not imply numerical invariance of risk. Relative to original-support calculations, the maximum BPI 2019 robust-risk change is 54.98 h on the 12 h grid and 33.53 h on the 24 h grid. The latter changes one saturation classification. The corresponding BPI 2017 maxima are 11.84 h and 25.54 h, with no classification changes. Table A4 reports these differences. The primary calculation does not depend on a grid.

3.7. Information Availability and Ranking Stability

Among BPI 2019 H1 origin events, 18.31% lack a successor observable at the decision cutoff, including events with no successor anywhere in the supplied trace. This proportion rises from 2.94% for January origins to 40.34% for June origins. The 90-day and 120-day

maturity windows reduce it to 6.76% and 4.32%. They produce 35 and 25 eligible candidates. Their original-DRO K = 5 sets remain identical to the primary set. BPI 2017 gives 34 and 27 mature-window candidates, also with the primary original-DRO top five. These comparisons do not remove uncertainty about the events that have no recorded successor.

At K = 5, BPI 2019 mean burden has a higher Kendall agreement with subsequent mean burden than original DRO, with 0.811 versus 0.733. Their eligible mean-oracle Jaccard overlaps are both 0.429. Empirical tail has the larger tail agreement, with 0.820 compared with 0.767 for original DRO. In BPI 2017, the four severity-based methods share a top-five set despite different full-ranking agreements. Their mean-oracle overlap is 0.667 and their tail-oracle overlap is 1.000. Table 8 summarizes historical-to-Q3 rank agreement and oracle overlap.

Table 8. Historical-to-Q3 Kendall tau-b and top-five overlap with eligible-candidate oracles. Mean and tail columns refer to separate burden targets.

Log	Method	Tau Mean	Tau Tail	J Mean Oracle	J Tail Oracle
2019	Frequency	0.395	0.462	0.250	0.429
2019	Mean burden	0.811	0.768	0.429	0.667
2019	Empirical tail	0.781	0.820	0.429	0.667
2019	Original DRO	0.733	0.767	0.429	0.667
2017	Frequency	0.087	0.123	0.111	0.111
2017	Mean burden	0.917	0.890	0.667	1.000
2017	Empirical tail	0.849	0.895	0.667	1.000
2017	Original DRO	0.859	0.880	0.667	1.000
2017	Saturation-aware DRO	0.859	0.880	0.667	1.000

The historical full-refit bootstrap gives BPI 2019 K = 5 mean Jaccard values of 0.911 for frequency and 1.000 for mean burden, empirical tail and original DRO. For BPI 2017, the values are 0.886, 0.897, 1.000, 0.981 and 0.992 for frequency, mean burden, empirical tail, original DRO and saturation-aware DRO. Candidate eligibility varies across the refits. Table A6 reports quantiles and eligible-set ranges, and the repository supplies transition-level inclusion probabilities.

4. Discussion

4.1. Severity Headroom and Selection Performance

BPI 2019 shows substantial saturation of worst-case tail severity, but not an incremental top-five loss caused by robustification. Original DRO and its zero-radius tail counterpart select the same five transitions. Their lower mean-burden capture relative to mean burden cannot therefore be attributed solely to the positive Wasserstein radius. The scores emphasize different outcomes. This distinction matters when mean burden is the operational objective.

The saturation ratio describes how much ambiguity consumes the remaining CVaR-to-cap headroom. It does not measure the number of distinguishable weighted scores, the gap at the selection boundary or future regret. In BPI 2019, all five selected robust risks equal the cap, yet unequal frequencies produce a substantial score gap at K = 5. At K = 10, a membership change is observed as the radius increases. The practical assessment must inspect the selected boundary, not only the fraction of saturated candidates.

4.2. Cross-Log Regime Interpretation

BPI 2017 has a much smaller median radius relative to its saturation threshold. Three induced saturations disappear under clipping, but the primary membership set remains identical across the four severity-based methods. The cross-log result demonstrates con-

trasting risk regimes. It does not demonstrate that a low ratio yields better prediction or that a high ratio necessarily invalidates a selected set.

Agreement between a threshold and a numerical risk calculation is an implementation check. A useful decision intervention would require evidence that acting on the diagnostic changes the selection in a beneficial way. That benefit is not established by the two primary K = 5 comparisons. A further independently specified window or process would be needed for a stronger decision-value claim.

4.3. Relation to Process-Mining and DRO Literature

The study differs from waiting-time root-cause analysis [2,3] because it does not claim that inter-event elapsed time equals a causally identified waiting mechanism. It differs from prescriptive process monitoring [15–19] because it does not estimate treatment effects or prescribe interventions for individual running cases. Instead, it addresses an upstream decision-support problem: which transitions should be selected for further investigation when future delay distributions may differ from the historical period. The direct-follow transition is therefore a prioritization unit, not a causal treatment target.

The model combines established Wasserstein ambiguity [9,10], CVaR [8] and bounded-loss risk structure [20]. Its application-specific contribution is the explicit connection between severity saturation, frequency-weighted transition scores and observed portfolio behavior. The exact expression removes the need for a discretized optimizer in the primary calculation. The transport and LP formulations remain useful for interpreting and checking the saturation mechanism.

4.4. Practical Implications

A process owner can use the threshold to identify transitions whose robust severity has reached the imposed cap. The next assessment is the score gap around the intended K and the agreement with mean-oriented and empirical-tail portfolios. Clipping may restore positive severity headroom, but only by narrowing the set of protected distributions. For an already empirically saturated transition, radius reduction cannot recover that headroom. Selected transitions remain candidates for investigation. Root-cause or intervention analysis is needed before inferring which operational change would reduce delay [2–19].

4.5. Limitations and Threats to Validity

The outcome is inter-event elapsed time, not an identified service or waiting component. Observed burden does not establish an intervention effect or account for intervention costs. The decision-snapshot rule prevents the use of later timestamped outcomes under the timestamp-as-availability assumption. It does not verify ingestion availability, and it conditions the historical sample on recorded completion. Maturity-lag checks reduce the share of later-recorded successors but cannot identify destinations missing from the log. Events without a recorded successor may represent either legitimate termination or incomplete follow-up. No inverse-probability correction is applied because the unobserved destination and completion mechanism are not identified.

Historical frequencies may drift, as may conditional delay distributions. The ambiguity set protects only the capped-delay distribution, so frequency changes can alter future burden shares even when the conditional risk model is adequate. Joint frequency-and-delay ambiguity would require a different formulation. The cap, tail level, calibration partition and minimum block size also affect saturation. Weekly resampling has few time blocks, while case-cluster resampling assumes independence between cases and does not preserve shared calendar shocks. Neither scheme identifies all process dependence. Two public logs and their Q3 comparisons also provide limited evidence for transfer to other processes.

5. Conclusions

Frequency-weighted Wasserstein–CVaR scores provide a bounded-risk view of transition prioritization. For one-dimensional absolute-distance ambiguity, robust CVaR equals empirical CVaR plus the radius divided by the tail probability, capped at C. The corresponding saturation threshold is the transport cost needed to fill that tail at C. The threshold and its normalized ratio describe severity headroom under the stated cap and calibration.

The available historical snapshots yield 33 saturated candidates out of 61 in BPI 2019 and three out of 41 in BPI 2017. At K = 5, original DRO and empirical tail have identical BPI 2019 membership. In BPI 2017, all four severity-based methods share the same membership, including the clipped-radius method. The primary comparisons show no incremental benefit from robustification or clipping. At larger K, membership changes can occur and must be evaluated separately.

Saturation of a risk component does not by itself establish deterioration in a frequency-weighted portfolio. The diagnostic is most informative when reported with completion availability, calibration sensitivity, selection-boundary gaps and future burden capture. Any reduction in the ambiguity radius should be accompanied by an explicit account of the protection it removes.

Supplementary Materials: The analytical repository version 1.1.0 contains source checksums, historical scores, selected portfolios, mean and tail evaluation tables, paired bootstrap comparisons, selection-refit outputs and sensitivity configurations. The analysis manifest links the exact code and aggregate-table checksums. Raw event logs are linked to their source repositories and are not redistributed. The versioned analysis code, checksum manifest and derived evaluation outputs are available at https://github.com/mpcristescu/bpi2019-robust-bottleneck-prioritization (accessed on 17 September 2026).

Funding: This research was funded by the Lucian Blaga University of Sibiu, grant number 6106/06.12.2024. The APC was funded by the Lucian Blaga University of Sibiu. The project was financed by the Lucian Blaga University of Sibiu through the research grant LBUS-IRG-2024.

Institutional Review Board Statement: Not applicable. The study analyzes publicly available, anonymized process-mining event logs.

Informed Consent Statement: Not applicable.

Data Availability Statement: The study uses two publicly available event logs. BPI Challenge 2019 is available from 4TU.ResearchData, DOI: 10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1 [13]. BPI Challenge 2017 is available from 4TU.ResearchData, DOI: 10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b [14]. The versioned analytical code and derived summary outputs are available in the public repository https://github.com/mpcristescu/bpi2019-robust-bottleneck-prioritization (accessed on 17 September 2026). The original event-log files are not redistributed. The availability-aware analysis is implemented in analysis/data.py, analysis/model.py, analysis/evaluation.py and analysis/run_analysis.py. Derived boundary and comparison tables are produced by analysis/publication_tables.py. The specification is version 1.1.0 and the primary random seed is 20260914. The manifest records source, code, portfolio and table SHA-256 checksums. The exported tables are under results/availability_analysis/bpi2019 and results/availability_analysis/bpi2017. The tests verify fractional empirical tails, the transport threshold, LP agreement, the destination cutoff and exact-zero comparisons for identical portfolios.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Bounded-Support Risk Characterization

Let P be a probability distribution supported on $[0, C]$, let q be its generalized inverse distribution function, and let $\beta = 1 - \alpha$. Write $b = \mathrm{CVaR}_{\alpha}(P)$. No moment or shape constraints are imposed on the ambiguity set. On the compact support, the Wasserstein ball is a compact

convex set of probability measures. The hinge expectation is continuous and affine in the measure and continuous and convex in t. The convex minimax interchange therefore identifies the robust functional in (5) with the supremum of CVaR over that ball.

For any supported Q with quantile function q_Q, the upper-tail quantile representation gives

$$\begin{aligned}\mathrm{CVaR}_\alpha(Q) - \mathrm{CVaR}_\alpha(P) &= \tfrac{1}{\beta}\textstyle\int_\alpha^1 q_Q(u) - q(u)\, du \\ &\le \tfrac{1}{\beta}\textstyle\int_0^1 \left|q_Q(u) - q(u)\right| du = \tfrac{W_1(P,Q)}{\beta}\end{aligned}$$

CVaR is also bounded above by C. Thus, robust CVaR cannot exceed the minimum of C and the affine upper bound. To attain that bound, modify only the upper β mass of the reference quantile. For a non-negative shift h, define

$$q_h(u) = \begin{cases} q(u) & 0 < u \le \alpha \\ \min\{C,\ q(u) + h\} & \alpha < u < 1 \end{cases}$$

The function q_h is non-decreasing and stays in $[0, \mathrm{C}]$, so it is a valid quantile function. Its Wasserstein transport cost is

$$d(h) = W_1(P, Q_h) = \int_\alpha^1 \min\{h,\ C - q(u)\}\, du$$

The cost $d(h)$ is continuous and non-decreasing. It starts at zero and reaches the following value by $h = C$.

$$d(C) = \int_\alpha^1 [C - q(u)]\, du = \beta(C - b)$$

If the budget is between zero and this endpoint, continuity supplies h with $d(h)$ equal to that budget. The modified distribution increases CVaR by exactly $\frac{d(h)}{\beta}$ and therefore attains the affine upper bound. If the budget reaches or exceeds the endpoint, $q_h = C$ on the entire upper tail and CVaR equals C. When the endpoint is zero, empirical CVaR already equals C and no positive transport is needed. Consequently,

$$\begin{aligned} R(\varepsilon) &= \min\left\{C, b + \tfrac{\varepsilon}{\beta}\right\} \\ \varepsilon_{\mathrm{sat}} = \beta(C - b),\ R(\varepsilon) = C &\ \Leftrightarrow\ \varepsilon \ge \varepsilon_{\mathrm{sat}} \end{aligned}$$

This proves the saturation equivalence and the closed-form threshold. In the discrete formulation (7), the cheapest source mass is the largest support below C, with an atom split if necessary. The transport cost is the integral of $C - q(u)$ over the part of the upper tail not already at C. That integral is exactly (8) and equals $\beta(C - b)$, including the case with sufficient empirical mass at C.

Appendix B. Availability and Sensitivity Results

Table A1. Origin-event availability. Later successors and absent recorded successors are retrospective audit categories, not inputs to historical selection.

Log	Window	Origin Events	Available Pairs	Later Successor	No Recorded Successor
2019	H1 snapshot	735,823	601,083	58,811	75,929
2019	H1, 90-day lag	314,535	293,273	3189	18,073
2019	H1, 120-day lag	184,825	176,843	1002	6980
2019	Q3, primary cutoff	418,716	352,493	0	66,223
2019	Q3, common 120-day age	367,086	312,457	0	54,629
2017	H1 snapshot	205,403	191,097	2097	12,209
2017	H1, 90-day lag	99,667	94,107	1	5559
2017	H1, 120-day lag	64,379	61,298	0	3081
2017	Q3, primary cutoff	138,217	129,191	0	9026
2017	Q3, common 120-day age	138,217	129,191	0	9026

No recorded successor is not a verified terminal-event label and is not automatically counted as right censoring. In BPI 2019, all recorded successors of Q3 origins occur before the primary observation cutoff. The 66,223 remaining Q3 origin events have no recorded successor in the supplied trace. BPI 2017 has 9026 such events. Assigning these events to a future transition would require unobserved information.

Table A2. Radius sensitivity. Overlap is relative to the primary original-DRO portfolio at the same K.

Log	Multiplier	Saturated *N*	J3	J5	J10	K10 Mean (%)
2019	0	3	1.000	1.000	0.818	88.688
2019	0.5	26	1.000	1.000	0.818	88.688
2019	1	33	1.000	1.000	1.000	87.430
2019	1.5	38	1.000	1.000	1.000	87.430
2019	2	42	1.000	1.000	1.000	87.430
2017	0	0	1.000	1.000	0.818	81.312
2017	0.5	1	1.000	1.000	0.818	82.820
2017	1	3	1.000	1.000	1.000	84.568
2017	1.5	3	1.000	1.000	1.000	84.568
2017	2	4	1.000	0.667	1.000	84.568

The boundary audit evaluates all saturation breakpoints in the multiplier interval $[0, 2]$. Between adjacent breakpoints every transition score is affine. For BPI 2019, the minimum selected-versus-unselected score margin remains positive for K = 3 and K = 5 at every breakpoint, which establishes membership invariance between them. For K = 10, the membership changes described in Section 3.3 are recorded in radius_boundary_audit.csv.

Table A3. Isolated clipping-factor sensitivity in BPI 2017. The original saturation total is three in every row.

Kappa	Clipped *N*	SA Saturated *N*	J5	Mean Share (%)	Tail Share (%)
0.50	4	0	1.000	70.237	60.461
0.75	3	0	1.000	70.237	60.461
0.90	3	0	1.000	70.237	60.461
0.95	3	0	1.000	70.237	60.461
0.99	3	0	1.000	70.237	60.461

Table A4. Grid effects relative to the original empirical support. Risk change is the maximum absolute change across primary candidates.

Log	Support	Saturated *N*	Median Ratio	Max Risk Change (h)	Changed Classes
2019	Original	33	1.13227	0.00	0
2019	12 h	33	1.13199	54.98	0
2019	24 h	34	1.15426	33.53	1
2017	Original	3	0.00615	0.00	0
2017	12 h	3	0.00603	11.84	0
2017	24 h	3	0.00658	25.54	0

Table A5. Paired 95% percentile intervals in percentage points, using 1000 replicates per scheme. The contrasts are original DRO minus mean burden for BPI 2019 and SA-DRO minus frequency for BPI 2017.

Log	Scheme	Clusters	Mean Interval	Tail Interval
2019	7-day blocks	14	[−5.99, −0.16]	[−0.42, 3.94]
2019	14-day blocks	7	[−5.53, −1.55]	[−0.02, 2.86]
2019	28-day blocks	4	[−7.36, −1.70]	[−0.46, 2.32]
2019	Case clusters	87,775	[−3.50, −3.01]	[1.51, 1.93]
2017	7-day blocks	14	[61.38, 64.44]	[40.50, 45.08]
2017	14-day blocks	7	[61.96, 64.00]	[40.36, 45.18]
2017	28-day blocks	4	[61.78, 64.27]	[40.49, 45.77]
2017	Case clusters	11,263	[62.15, 63.61]	[41.28, 44.16]

For identical membership sets, the mean and tail differences remain exactly zero under all four dependence schemes. In BPI 2019, the mean disadvantage relative to mean burden persists across these schemes. The tail advantage has intervals containing zero under the calendar schemes, but not under case-cluster resampling. This dependence sensitivity limits a significance claim for the tail comparison.

Table A6. Historical top-five selection stability from 500 full block refits. Candidate range is the eligible count across the same refits.

Log	Method	Mean J	5th Percentile	95th Percentile	Candidate Range
2019	Frequency	0.911	0.667	1.000	46 to 67
2019	Mean burden	1.000	1.000	1.000	46 to 67
2019	Empirical tail	1.000	1.000	1.000	46 to 67
2019	Original DRO	1.000	1.000	1.000	46 to 67
2017	Frequency	0.886	0.667	1.000	39 to 43
2017	Mean burden	0.897	0.667	1.000	39 to 43
2017	Empirical tail	1.000	1.000	1.000	39 to 43
2017	Original DRO	0.981	0.667	1.000	39 to 43
2017	Saturation-aware DRO	0.992	1.000	1.000	39 to 43

Table A7. Identical primary top-five membership with method-specific ranks in BPI 2019.

Transition	Tail Rank	DRO Rank
Record Invoice Receipt to Clear Invoice	1	2
Create Purchase Order Item to Vendor Creates Invoice	2	1
Record Goods Receipt to Record Invoice Receipt	3	3
Vendor Creates Invoice to Record Invoice Receipt	4	4
Create Purchase Order Item to Record Goods Receipt	5	5

Table A8. Identical primary top-five membership with method-specific ranks in BPI 2017.

Transition	Mean Rank	Tail Rank	DRO Rank	SA Rank
A_Complete to A_Cancelled	1	3	3	3
A_Complete to A_Validating	2	1	1	1
A_Concept to A_Accepted	3	2	2	2
A_Complete to O_Create Offer	4	5	4	4
A_Incomplete to A_Validating	5	4	5	5

Table A9. Aggregate-output mapping for version 1.1.0. Paths are relative to the respective log directory under results/availability_analysis.

Manuscript Material	Aggregate Output
Tables 3 and 5	holdout_evaluation.csv
Table 4	historical_scores.csv and summary.json
Tables 6 and 7	sensitivity_summary.csv
Table 8	holdout_evaluation.csv
Table A1	availability_audit.csv
Table A2	sensitivity_summary.csv and radius_boundary_audit.csv
Table A3	clipping_diagnostics.csv and sensitivity_summary.csv
Table A4	grid_comparison.csv and sensitivity_transition_risks.csv
Table A5	paired_bootstrap.csv
Table A6	selection_stability.csv and selection_probabilities.csv
Tables A7 and A8	portfolios.csv
Mathematical checks	mathematical_checks.csv
Sampling benchmark	sampling_variability_benchmark.csv

References

1. van der Aalst, W.M.P. *Process Mining: Data Science in Action*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2016. [CrossRef]
2. Lashkevich, K.; Milani, F.; Chapela-Campa, D.; Suvorau, I.; Dumas, M. Unveiling the causes of waiting time in business processes from event logs. *Inf. Syst.* **2024**, *126*, 102434. [CrossRef]
3. Tang, J.; Liu, Y.; Lin, K.-Y.; Li, L. Process bottlenecks identification and its root cause analysis using fusion-based clustering and knowledge graph. *Adv. Eng. Inform.* **2023**, *55*, 101862. [CrossRef]
4. Sato, D.M.V.; de Freitas, S.C.; Barddal, J.P.; Scalabrin, E.E. A Survey on Concept Drift in Process Mining. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
5. Kraus, A.; van der Aa, H. Machine learning-based detection of concept drift in business processes. *Process Sci.* **2025**, *2*, 5. [CrossRef]
6. Kraus, A.; van der Aa, H. Comprehensive characterization of concept drifts in process mining. *Inf. Syst.* **2026**, *135*, 102584. [CrossRef]
7. Hinder, F.; Vaquet, V.; Hammer, B. One or two things we know about concept drift—A survey on monitoring in evolving environments. Part A: Detecting concept drift. *Front. Artif. Intell.* **2024**, *7*, 1330257. [CrossRef] [PubMed]
8. Rockafellar, R.T.; Uryasev, S. Optimization of conditional value-at-risk. *J. Risk* **2000**, *2*, 21–41. [CrossRef]
9. Mohajerin Esfahani, P.; Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **2018**, *171*, 115–166. [CrossRef]
10. Blanchet, J.; Murthy, K. Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* **2019**, *44*, 565–600. [CrossRef]
11. Leemans, S.J.J.; van der Aalst, W.M.P.; Brockhoff, T.; Polyvyanyy, A. Stochastic process mining: Earth movers' stochastic conformance. *Inf. Syst.* **2021**, *102*, 101724. [CrossRef]
12. Leemans, S.J.J.; Polyvyanyy, A. Stochastic-aware precision and recall measures for conformance checking in process mining. *Inf. Syst.* **2023**, *115*, 102197. [CrossRef]
13. van Dongen, B. BPI Challenge 2019. 4TU.ResearchData, 2019. Dataset. Available online: https://research.tue.nl/nl/datasets/bpi-challenge-2019/ (accessed on 17 September 2026).
14. van Dongen, B. BPI Challenge 2017. 4TU.ResearchData, 2017. Dataset. Available online: https://www.kaggle.com/datasets/ras-canudragos/bpi-challenge-2017 (accessed on 17 September 2026).
15. Kubrak, K.; Milani, F.; Nolte, A.; Dumas, M. Prescriptive process monitoring: Quo vadis? *PeerJ Comput. Sci.* **2022**, *8*, e1097. [CrossRef] [PubMed]
16. Dasht Bozorgi, Z.; Teinemaa, I.; Dumas, M.; La Rosa, M.; Polyvyanyy, A. Prescriptive process monitoring based on causal effect estimation. *Inf. Syst.* **2023**, *116*, 102198. [CrossRef]
17. Shoush, M.; Dumas, M. Prescriptive Process Monitoring Under Resource Constraints: A Causal Inference Approach. In *Process Mining Workshops, ICPM 2021*; LNBIP 433; Springer: Cham, Switzerland, 2022; pp. 180–193. [CrossRef]
18. Shoush, M.; Dumas, M. White box specification of intervention policies for prescriptive process monitoring. *Data Knowl. Eng.* **2025**, *155*, 102379. [CrossRef]

19. Hundogan, O.A.; Verhoef, B.J.; Theeven, P.; Reijers, H.A.; Lu, X. Reinforcement learning for optimizing responses in care processes. *Data Knowl. Eng.* **2025**, *157*, 102412. [CrossRef]
20. Cai, J.; Liu, F.; Yin, M. Worst-case distortion risk measures of transformed losses with uncertain distributions lying in Wasserstein balls. *ASTIN Bull.* **2026**, *56*, 270–294. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.