

Queue Shock at Model Upgrades: A Longitudinal Audit of EPSS v2–v5 Rank Stability and Prioritization Workload

Cybersecurity Advances • Volume 3 • Issue 1 • 2026 • pp. 1–13

ATS PRESS Research Group

AlphaTech Scholarly Press, United States

Correspondence: Md Sazzad Hossain • md@hossain.tech • ORCID 0009-0009-4948-1179

Received 2026-10-01 • Accepted 2026-10-02 • Published 2026-10-02

© 2026 ATS PRESS Research Group. Published by AlphaTech Scholarly Press under CC BY 4.0.

Abstract

Background: EPSS publishes daily probabilities and percentiles to help defenders prioritize vulnerabilities. Major model replacements may disrupt operational queues, but this discontinuity has not been quantified across public EPSS versions.

Objective: We measured how EPSS v2–v3, v3–v4, and v4–v5 transitions changed rankings, bounded priority sets, threshold classifications, and score coverage.

Methods: A protocol was frozen before primary execution. We acquired 96 official daily archives covering 30 adjacent-day controls before each transition, the preceding day, and the transition day. Analyses used CVEs present on both compared days and reported additions and removals separately. Prespecified outcomes included rank correlations, score and percentile displacement, Jaccard similarity for five queue definitions, six threshold policies, and time-aligned CISA KEV stratification. The tested pipeline was rerun to verify a canonical result hash.

Results: Transition-day Spearman correlations were 0.670, 0.749, and 0.834, compared with control medians of 0.99994, 0.99988, and 0.99962. Top-1,000 Jaccard similarities were 0.321, 0.280, and 0.416. The 95th percentile of absolute percentile movement reached 0.476, 0.437, and 0.352. At EPSS ≥ 0.10 , 4.79%, 5.36%, and 4.02% of shared CVEs changed classification. The v3–v4 transition removed 11,576 CVEs and added 59. Every primary transition shock exceeded all 30 matched controls.

Conclusions: A major EPSS release is a prioritization regime change, not an ordinary daily refresh. Defenders should version score-based policies, dual-run queues, budget for reclassification, and retain pre-transition snapshots.

Keywords: EPSS; vulnerability prioritization; model versioning; rank stability; remediation queue; CISA KEV; cybersecurity operations

1. Introduction

Vulnerability management is a constrained allocation problem. Published CVEs accumulate faster than most organizations can test and remediate them, so defenders need a defensible way to decide what enters the next patch cycle. The Exploit Prediction Scoring System (EPSS) addresses one part of that problem by estimating the probability that exploitation activity will be observed for a CVE within the next 30 days [1,2]. FIRST publishes scores and percentiles daily, allowing teams to rank vulnerabilities or apply score-based policies [3–5].

A daily feed can create the impression of continuity. That assumption is reasonable when the same model ingests one additional day of evidence. It is less safe when the scoring function itself changes. FIRST has released four successive model families in the period covered by its public archive: v2, v3, v4, and v5 [3].

Each major version can change features, training data, model form, calibration, or coverage. Even if a replacement improves predictive performance, it can still reorder an operational queue abruptly.

Queue stability matters because remediation is not instantaneous. Tickets are assigned, exceptions are approved, maintenance windows are booked, and service owners are contacted before a patch is deployed. A model transition that replaces half of a top-priority set may invalidate ongoing work. Fixed score thresholds can create a different burden: thousands of CVEs can cross an escalation boundary overnight even when no corresponding change occurs in an organization's assets. Percentile thresholds avoid some calibration drift, but they can still reassign membership when ranks change.

Prior work established the motivation and empirical design of EPSS [1], evaluated later model improvements [2], and studied how scores evolve after vulnerability disclosure [7]. Other work has examined EPSS against known exploitation, current-versus-historical scoring, or patching workload. These questions concern predictive validity or temporal behavior within a model. They do not quantify the operational discontinuity at all three public major-version boundaries from v2 through v5.

We address that gap with a protocol-locked longitudinal audit of official daily EPSS archives. The study asks five questions. First, how much do transition days change scores, percentiles, and ranks relative to ordinary adjacent days? Second, how stable are fixed-capacity priority sets? Third, how many CVEs cross common score and percentile thresholds? Fourth, are vulnerabilities already present in the CISA KEV catalog more stable? Fifth, how much of the observed change reflects rank reassignment versus additions to or removals from the scored universe?

The contribution is operational rather than predictive. We do not test whether one EPSS version forecasts exploitation better than another. We measure the migration burden created when a production prioritization model changes. The resulting evidence supports version-aware runbooks, dual-run deployment, and auditable change control.

2. Background and related work

2.1 EPSS as a prioritization signal

The original EPSS work framed exploitation prediction as a way to move beyond severity-only remediation strategies [1]. Later work expanded the data and modeling approach and evaluated the tradeoff between coverage of exploited vulnerabilities and remediation effort [2]. EPSS probabilities and percentiles serve different operational purposes. A probability estimates an absolute event likelihood under the model; a percentile describes relative position among scored CVEs [4,5,9].

This distinction matters at a model boundary. A calibration change can move many absolute scores while leaving relative order less affected. A ranking change can disrupt percentile-based and fixed-capacity queues even when absolute score distributions appear stable. Assessing only mean or median scores therefore misses the action taken by a triage system.

2.2 Temporal behavior and exploitation evidence

Ravalico and colleagues examined longitudinal EPSS trajectories for newly disclosed vulnerabilities [7]. Their work establishes that scores evolve over time and that early estimates can differ from later ones. Retrospective performance studies also face look-ahead bias when they apply current scores to historical decisions. Those observations motivate date-matched data, but our estimand differs: we compare the same shared CVEs immediately before and after a major model replacement, then benchmark that discontinuity against the preceding 30 ordinary daily comparisons.

CISA's KEV catalog provides evidence that listed vulnerabilities have been exploited in the wild and imposes remediation requirements on US federal civilian agencies [6]. KEV is not a complete or

instantaneous label for all exploitation, so we do not use it as predictive ground truth. We use `dateAdded` only to define a time-aligned subgroup already known to be exploited before each comparison. This avoids back-projecting later catalog knowledge into earlier dates. NIST's work on likely exploited vulnerabilities likewise cautions that remediation signals answer different questions and should be interpreted in their intended decision context [8].

2.3 Model replacement as an operational event

Machine-learning governance often separates routine inference from model release, yet vulnerability workflows frequently ingest a single rolling feed without retaining its model identifier. That design hides a change in policy semantics. A ticket created under v3 may no longer qualify under v4, while an audit performed later may retrieve only the newer score. The public EPSS archive includes model-version metadata, making it possible to measure these discontinuities directly [3].

We use three complementary views. Rank correlations measure global ordering. Set similarity measures whether bounded patch queues retain the same members. Threshold crossings measure policy reclassification. Coverage changes identify CVEs that enter or leave the scored universe and therefore cannot be evaluated through shared-record deltas alone.

3. Methods

3.1 Protocol and study design

The protocol was frozen and hashed before the primary transition corpus was acquired or outcome metrics were calculated. Its SHA-256 digest is

16f1b41142aada0246f32670020bef3953a08739d9fec718176514e2d488a83f. No outcome-driven protocol amendments were made.

We studied the transition dates documented by FIRST [3]: 7 March 2023 for v3, 17 March 2025 for v4, and 15 June 2026 for v5. For each event, the primary comparison joined the archive from the preceding calendar day to the first archive under the new version. The control distribution comprised 30 consecutive adjacent-day comparisons ending before the transition pair. This required 32 snapshots per event and 96 snapshots in total.

The design uses a local discontinuity comparison rather than a causal estimate. Ordinary controls characterize the scale of movement under the outgoing model immediately before replacement. They do not prove that every transition-day difference was caused solely by the model code, because new evidence and coverage changes can arrive on the same day. The magnitude relative to the preceding daily envelope nevertheless identifies how unusual the transition was for a feed consumer.

3.2 Data acquisition and integrity

The pipeline downloaded dated gzip archives from the official EPSS archive endpoint. Each file contains a comment header with model version and score timestamp followed by `cve`, `epss`, and `percentile` columns. The parser failed closed on a missing header, date mismatch, duplicate CVE, absent field, nonnumeric value, or value outside [0,1]. A manifest records source URL, byte size, SHA-256 digest, row count, model version, and available HTTP metadata.

We acquired the official CISA KEV JSON feed on 1 October 2026. It contained 1,730 unique CVEs. For a comparison beginning on date d , the KEV subgroup included only records with `dateAdded` $\leq d$. Future catalog additions were excluded.

3.3 Cohorts and coverage

Score and rank analyses used the intersection of CVEs present on both days. This shared-CVE restriction prevents records that enter or leave the feed from being assigned artificial score changes. We reported additions, removals, and coverage Jaccard similarity separately.

The shared universe contained 196,815 CVEs for $v2 \rightarrow v3$, 270,505 for $v3 \rightarrow v4$, and 340,204 for $v4 \rightarrow v5$. Time-aligned KEV subsets contained 884, 1,302, and 1,619 CVEs, respectively.

3.4 Rank and movement measures

For each pair, we calculated Spearman correlation over EPSS scores and Kendall tau-b, which accounts for ties. We also calculated absolute changes in score, official percentile, and normalized rank. Average ranks were used for tied scores. Normalized rank was $(\text{rank}-1)/(n-1)$, where zero denotes the highest priority.

Movement distributions were summarized by median, 90th, 95th, and 99th percentiles. These summaries are population descriptions of the archived score universe, not estimates from a random sample. We therefore emphasize effect sizes and control envelopes rather than null-hypothesis tests.

3.5 Fixed-capacity priority sets

We evaluated top 100, top 1,000, top 10,000, top 1%, and top 5% sets. Ordering was deterministic: descending score, descending official percentile, then ascending CVE identifier. For each set, we reported intersection, union, additions, removals, retention, and Jaccard similarity. These sets approximate a bounded global work queue. They do not represent any particular organization's installed software inventory.

3.6 Threshold policies

We evaluated score thresholds of 0.01, 0.10, and 0.50 and percentile thresholds of 0.90, 0.95, and 0.99. For each threshold, the pipeline counted upward crossings, downward crossings, total changed classifications, and queue sizes before and after. Thresholds were chosen before the primary run to cover permissive, moderate, and high score cutoffs and commonly interpreted upper percentile bands. FIRST does not prescribe one universal operational threshold [5].

3.7 Transition shock and sensitivity

Each primary endpoint was compared with the 30 ordinary controls for the same outgoing model: Spearman correlation, 95th-percentile absolute percentile movement, top-1,000 Jaccard similarity, and changed fraction at $\text{EPSS} \geq 0.10$. For movement metrics, larger values indicate greater shock; for stability metrics, smaller values do. We report the control median, range, ratio on the corresponding shock scale, and empirical position among the controls.

Seven-day and 14-day windows were prespecified as sensitivity views. Because all transition values exceeded the shock of every individual comparison in the full 30-day window, the shorter windows did not alter the direction or extremity of any primary finding.

3.8 Software verification

Core parsing, time alignment, pair construction, deterministic ranking, set comparison, and threshold logic were developed with test-first checks. The final suite contained 12 tests. The complete study was executed twice from the frozen local archives. Both runs produced the same canonical semantic-result SHA-256 digest: 40eb0446d2e0cb4411efcd4131cb8d3b491144d68f41672656fd4aa0d1c5ffcb.

4. Results

4.1 Coverage changed at every transition

Table 1 separates feed coverage from shared-CVE reordering. The v2→v3 boundary added 140 CVEs and removed one. The v3→v4 boundary had the largest contraction: 11,576 removals and 59 additions, reducing coverage Jaccard similarity to 0.959. The v4→v5 boundary removed 1,146 CVEs and added 43.

Table 1. Coverage and global rank stability across major EPSS transitions

Transition	Before CVEs	After CVEs	Shared	Added	Removed	Coverage Jaccard	Spearman	Kendall tau-b
v2→v3	196,816	196,955	196,815	140	1	0.999	0.670	0.510
v3→v4	282,081	270,564	270,505	59	11,576	0.959	0.749	0.568
v4→v5	341,350	340,247	340,204	43	1,146	0.997	0.834	0.651

A consumer that records only scores would miss the distinction between reordering and dropped coverage. The v3→v4 transition is especially important: a large set of previously scored CVEs disappeared while the shared records were also substantially reordered.

4.2 Transition-day ordering was unlike ordinary daily movement

Spearman correlations fell to 0.670, 0.749, and 0.834 at the three boundaries (Figure 1). The corresponding 30-day control medians were 0.999941, 0.999876, and 0.999620. Even the least stable control day had correlations of 0.9950, 0.9993, and 0.9990. Thus, each transition was separated from its entire local control range.

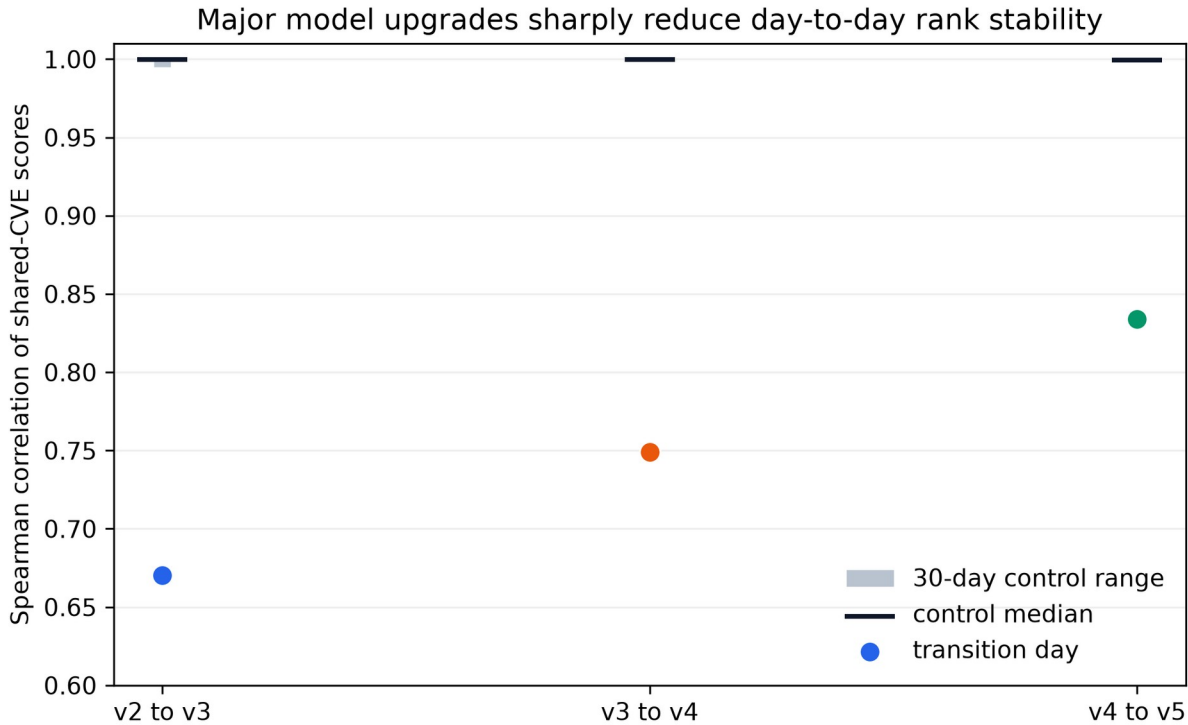


Figure 1. Spearman score correlation on each transition day compared with the median and range of the 30 preceding ordinary adjacent-day comparisons.

Percentile displacement was broad. Median absolute percentile changes were 0.1198 for v2→v3, 0.1079 for v3→v4, and 0.0847 for v4→v5. At the 95th percentile, movement reached 0.4763, 0.4370, and 0.3517. The matched control medians for the same 95th-percentile statistic were 0.00016, 0.00015, and 0.00046. Figure 4 shows that the effect was not confined to a small tail.

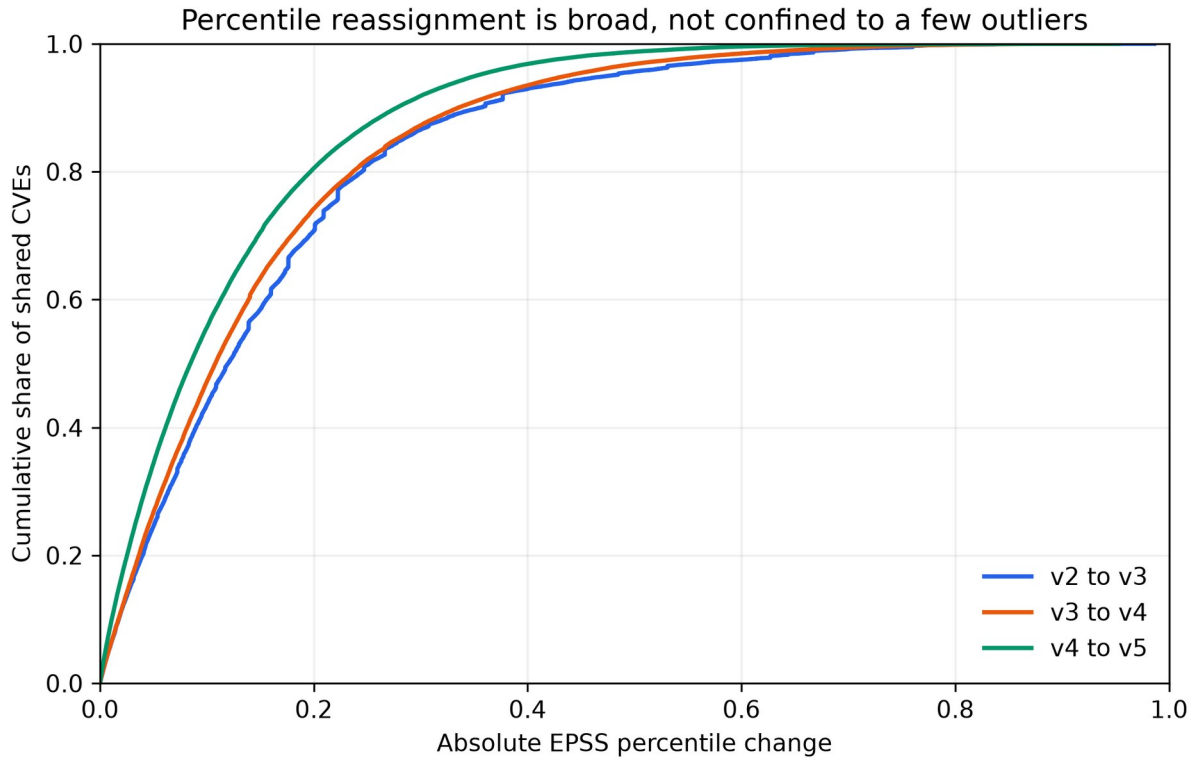


Figure 4. Empirical cumulative distributions of absolute official-percentile movement for shared CVEs at each model transition.

Absolute score movement and percentile movement conveyed different information. Median score changes were 0.00878, 0.00158, and 0.00446, while the associated percentile changes were roughly one-tenth of the full percentile scale. A small probability change can therefore represent a large change in relative priority, particularly where the score distribution is dense.

4.3 Fixed-capacity queues experienced high turnover

Table 2 reports set stability. For the top 1,000, only 486, 438, and 588 members remained after the v3, v4, and v5 upgrades. The resulting Jaccard similarities were 0.321, 0.280, and 0.416. At top 100, retention was lower: 39%, 34%, and 37%.

Table 2. Fixed-capacity and fractional priority-set stability

Transition	Set	Retained	Added	Removed	Jaccard
v2→v3	Top 100	39	61	61	0.242
v2→v3	Top 1,000	486	514	514	0.321
v2→v3	Top 10,000	5,868	4,132	4,132	0.415
v3→v4	Top 100	34	66	66	0.205
v3→v4	Top 1,000	438	562	562	0.280
v3→v4	Top 10,000	5,699	4,301	4,301	0.399
v4→v5	Top 100	37	63	63	0.227
v4→v5	Top 1,000	588	412	412	0.416
v4→v5	Top 10,000	6,747	3,253	3,253	0.509

Fractional queues showed the same pattern (Figure 2). Top-1% Jaccard similarity ranged from 0.309 to 0.450; top-5% similarity ranged from 0.409 to 0.523. Larger queues absorbed more rank movement, but no major version preserved even 70% of top-5% membership.

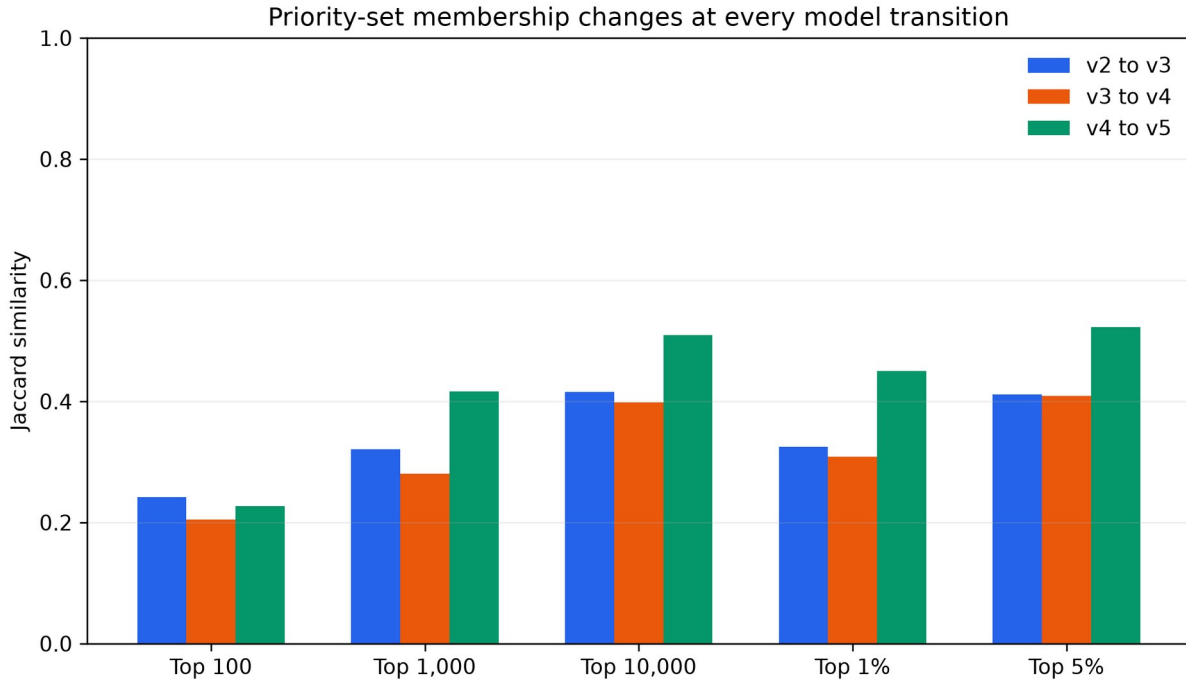


Figure 2. Jaccard similarity of five fixed-capacity or fractional shared-CVE priority sets across major model transitions.

Ordinary-day top-1,000 Jaccard medians were 1.000 for the v2 control window and 0.998 for the v3 and v4 windows. Each transition again represented the most disruptive pair in its 30-day local comparison.

4.4 Fixed score policies created asymmetric workload changes

Threshold results reveal changes that set similarity alone cannot show. At $EPSS \geq 0.10$, 9,419 shared CVEs changed classification under v3, 14,495 under v4, and 13,667 under v5. These counts correspond to 4.79%, 5.36%, and 4.02% of the shared universes.

The direction differed by model. At v2–v3, 5,333 CVEs moved above 0.10 and 4,086 moved below it, increasing the queue. At v3–v4, 10,428 moved above and 4,067 below. At v4–v5, only 3,820 moved above while 9,847 moved below, shrinking the ≥ 0.10 queue from 23,068 to 17,041.

Lower score thresholds were more volatile. At 0.01, v2–v3 reclassified 44.69% of shared CVEs, mostly downward; v3–v4 reclassified 14.01%; v4–v5 reclassified 22.88%, mostly upward. The threshold therefore did not carry a stable workload meaning across versions.

Percentile cutoffs limited net queue-size drift because they preserve approximate proportions, but membership still changed. At the 90th percentile, 7.09%, 8.23%, and 5.81% of shared CVEs crossed the boundary. At the 99th percentile, 1.02%, 1.08%, and 0.76% changed classification. Figure 3 compares the six policies.

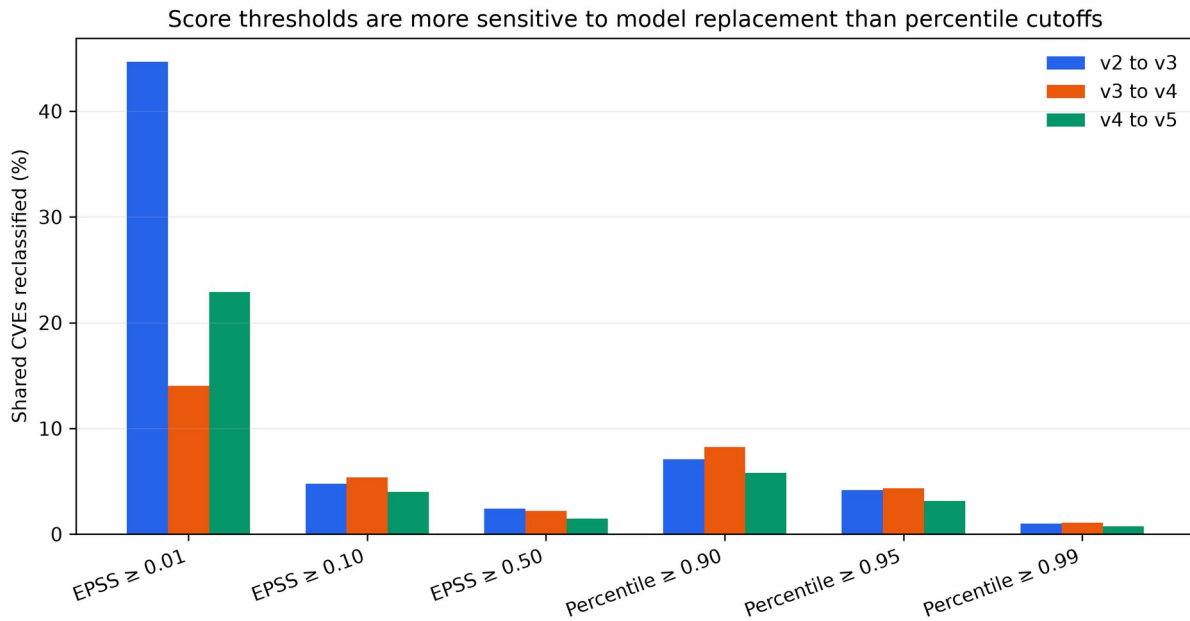


Figure 3. Fraction of shared CVEs crossing three score thresholds and three percentile thresholds at each model transition.

4.5 Already-known-exploited vulnerabilities were more stable, but not stable enough to ignore migration

The KEV-as-of subsets had higher Spearman correlations than the full universes: 0.726, 0.813, and 0.881. Median absolute percentile movement was 0.0512, 0.0177, and 0.0061, lower than the full-universe medians. This result is consistent with prominent exploitation evidence anchoring some high-risk CVEs.

The tail still moved substantially. The 95th percentiles of absolute percentile change among already-KEV records were 0.584, 0.427, and 0.265. A KEV designation also carries an independent action rationale, so an organization should not remove a known-exploited vulnerability from remediation solely because a new EPSS version lowers its score. Our subgroup analysis supports combining signals rather than treating a model migration as a reason to overwrite prior evidence.

4.6 All primary shocks exceeded the matched control envelope

For all 12 event-endpoint combinations, the transition was more disruptive than every one of the 30 controls. The empirical shock position was therefore 1.0 under the prespecified finite-sample ranking. The ratios were also large. The v2→v3, v3→v4, and v4→v5 95th-percentile movement was 2,977, 2,913, and 764 times the respective control median. Changed fractions at EPSS ≥ 0.10 were 3,132, 1,506, and 522 times the control median.

Ratios become unstable when ordinary-day movement is nearly zero, so these values should not be read as performance gains or causal multipliers. Their role is descriptive: transition days were orders of magnitude outside routine feed variation.

5. Discussion

5.1 Major releases should be deployed as migrations

The central result is simple: replacing an EPSS model is not operationally equivalent to loading the next daily file. Across all three upgrades, global rank correlation dropped sharply, 41% to 56% of top-1,000 entries were replaced, and thousands of CVEs crossed moderate score thresholds. A pipeline that treats the model identifier as incidental metadata conceals a change in decision policy.

A safer process has four stages. First, pin and record the model version beside every score used in a decision. Second, dual-run the outgoing and incoming versions on the organization's actual asset inventory. Third, calculate ticket additions, removals, and priority inversions before changing service-level agreements. Fourth, preserve the old snapshot and migration report so later auditors can reconstruct why a ticket was opened or closed.

This process does not reject model improvement. Predictive systems need revision as threat data and exploitation behavior change. The operational point is that a better model can still create transition cost, and that cost should be measured rather than discovered through a flooded ticket queue.

5.2 Score thresholds are version-specific policy parameters

A numerical threshold such as 0.10 looks portable because every version outputs a value between zero and one. Our results show that portability is unsafe. Queue size rose under v3 and v4 but fell under v5, with thousands of bidirectional crossings in each case. The same numeric cutoff therefore encoded different workloads across versions.

Percentiles gave more predictable queue sizes, but they did not preserve membership. For capacity planning, a percentile or fixed top-K policy may be easier to govern than an absolute score threshold. For risk estimation, the probability retains its substantive meaning, subject to calibration. Organizations need to decide which objective controls the policy and revalidate it at each model release.

Hybrid rules can reduce abrupt reversals. A team might retain already-assigned work for a grace period, require two consecutive days below a removal boundary, or combine EPSS with KEV status, asset exposure, reachability, business impact, and compensating controls. Such rules should be tested against local data. This study does not establish the best production policy.

5.3 Coverage needs its own monitoring

The v3–v4 transition removed 11,576 records from the scored universe, far more than were added. A missing score is not a low score. Consumers should monitor feed additions and removals, keep an explicit unknown state, and avoid silently dropping vulnerabilities from dashboards when a new archive no longer contains them.

Coverage Jaccard similarity can be monitored alongside row count. A sudden contraction should trigger investigation before downstream automation clears tickets. The same rule applies to API integrations that expose only current values: local archival copies are necessary to distinguish a removed record from one that was never present.

5.4 Implications for research and benchmarking

Historical EPSS studies should preserve model versions and use date-matched archives. Pooling scores across a major boundary without a version indicator combines different scoring regimes. A time-series discontinuity may reflect model replacement rather than new exploitation evidence. Researchers should report transition dates, conduct version-stratified sensitivity analyses, and avoid interpolating through those boundaries as if the data-generating process were unchanged.

Our results complement temporal trajectory work [7]. Daily evolution within a version and discontinuity between versions are separate phenomena. The former measures responsiveness; the latter measures migration shock. Both matter when interpreting historical score series.

6. Limitations

The analysis covers three public major transitions. Three events are sufficient for a complete audit of v2 through v5, but not for estimating a general statistical distribution of future upgrades. EPSS v1 predates the consistent archive metadata needed for the same design and was outside the frozen protocol.

The control window is local and descriptive. Ordinary days may include weekends, new disclosures, exploitation reports, and feed maintenance. We retained calendar-day comparisons because a consumer receives the feed on those days. Seven-day and 14-day sensitivity windows preserved the findings, but a different season or threat environment could have different routine volatility.

Shared-CVE analysis deliberately excludes additions and removals from movement statistics. Coverage is reported separately, but we cannot infer why individual CVEs entered or left the feed. Nor can we separate every component of a model transition, such as feature changes, retraining data, calibration, and eligibility rules, from public outputs alone.

The priority sets are global proxies. Real organizations see only vulnerabilities present in their assets and may enrich them with exposure, reachability, business value, and controls. Local turnover could be smaller or larger. The reported counts should support migration planning, not serve as a universal estimate of ticket volume.

CISA KEV is a curated catalog with its own scope and publication delay. We used it only for a time-aligned subgroup, not as complete exploitation ground truth. A lower EPSS score does not negate a KEV designation or other direct evidence.

Finally, this study measures stability rather than predictive quality. High stability can preserve obsolete priorities; low stability can reflect a justified correction. No result here ranks EPSS versions by accuracy or recommends retaining an older model indefinitely.

7. Reproducibility and data availability

The submission supplement contains the frozen protocol and checksum, test suite, acquisition and analysis code, download manifest, derived result tables, figures, environment lock, and file hashes. Raw EPSS archives are redistributable under the source's stated terms but are omitted from the compact journal upload if the portal imposes file-size limits; the manifest provides exact dated URLs and SHA-256 hashes for reacquisition. The CISA KEV snapshot and its hash are included where licensing and package limits permit.

The pipeline runs from official dated archives and fails on malformed metadata, duplicate identifiers, out-of-range values, or date mismatches. A second complete run from the frozen files reproduced the semantic result hash exactly.

8. Conclusion

Every major EPSS transition from v2 through v5 produced a discontinuity far outside the preceding month of daily variation. Global rank correlations fell to 0.670–0.834, top-1,000 Jaccard similarity fell to 0.280–0.416, and 4.02%–5.36% of shared CVEs crossed the 0.10 score threshold. Coverage also changed, most visibly when v4 replaced v3.

These findings do not argue against EPSS updates. They show that an update changes the operational meaning of a score-based queue. Vulnerability programs should treat each major release as a controlled migration: preserve versions, compare queues on local assets, monitor coverage, budget for reclassification, and keep direct exploitation evidence such as KEV separate from model-derived priority.

Declarations

Ethics statement

The study used public, non-personal cybersecurity datasets and performed no interaction with human participants, private systems, or live targets. Institutional review was not applicable.

Funding

No external funding was received for this study.

Competing interests

The collective author is affiliated with the publisher of the target journal. The author requests handling by an editor with no involvement in the study or publisher management and independent external peer review. The author will not participate in editorial selection, review, or acceptance decisions for this manuscript.

Author contributions

The collective author conceived the study, approved the protocol, supervised the computational work, interpreted the results, and accepts accountability for the submitted manuscript and reproducibility package. AI systems do not qualify for authorship or CRediT attribution.

Use of generative AI

Generative AI assisted with protocol development, literature organization, code and test generation, execution support, data analysis, visualization code, quality checks, and substantive manuscript drafting. The accountable human directing the collective authorized the study and submission and retains responsibility for verifying the evidence, approving the final text, and responding to editorial questions. No AI system is listed as an author.

Data and code availability

The dated public source URLs, hashes, code, tests, derived results, and figures are included in the reproducibility supplement submitted with the manuscript. The primary semantic result digest is 40eb0446d2e0cb4411efcd4131cb8d3b491144d68f41672656fd4aa0d1c5ffcb.

References

1. Jacobs J, Romanosky S, Edwards B, Adjerid I, Roytman M. Exploit Prediction Scoring System (EPSS). *Digital Threats: Research and Practice*. 2021;2(3):1–17. <https://doi.org/10.1145/3436242>
2. Jacobs J, Romanosky S, Suci O, Edwards B, Sarabi A. Enhancing Vulnerability Prioritization: Data-Driven Exploit Predictions with Community-Driven Insights. In: *2023 IEEE European Symposium on Security and Privacy Workshops (EuroSPW)*. 2023:194–206. <https://doi.org/10.1109/EuroSPW59978.2023.00027>
3. Forum of Incident Response and Security Teams. Get the Data: Exploit Prediction Scoring System. <https://www.first.org/epss/data>. Accessed 1 October 2026.
4. Forum of Incident Response and Security Teams. Exploit Prediction Scoring System. <https://www.first.org/epss/>. Accessed 1 October 2026.
5. Forum of Incident Response and Security Teams. Using EPSS. <https://www.first.org/epss/using-epss>. Accessed 1 October 2026.
6. Cybersecurity and Infrastructure Security Agency. Known Exploited Vulnerabilities Catalog. <https://www.cisa.gov/known-exploited-vulnerabilities-catalog>. Accessed 1 October 2026.

7. Ravalico D, Farina M, Trevisan M, Bartoli A. Analysing the Temporal Dynamics of the Exploit Prediction Scoring Systems (EPSS). 2025. <https://doi.org/10.2139/ssrn.5147459>
8. Mell P, Spring J. Likely Exploited Vulnerabilities. NIST Cybersecurity White Paper 41. 2025. <https://doi.org/10.6028/NIST.CSWP.41>
9. Forum of Incident Response and Security Teams. EPSS Frequently Asked Questions. <https://www.first.org/epss/faq>. Accessed 1 October 2026.