THE COMPUTERTECH
(An International Peer Review Journal)
VOLUME 12; ISSUE 1 (JAN-JUNE); (2026)
WEBSITE: THE COMPUTERTECH

Detecting and Correcting Temporal Drift in Credit-Risk and Fraud-Detection Models at Community Banks, Credit Unions, and CDFIs: An Automated Model-Risk-Monitoring Framework Below the Supervisory Threshold

Saeed Ur Rashid[1*], Shames Al Muzakkiruddin [2], Charles Møller[2]

[1]Westcliff University, California, USA
[2]University of Arkansas at Little Rock (ERIQ), USA
*Corresponding Author Email: labib668@gmail.com

Abstract

On April 17, 2026, the Federal Reserve, OCC, and FDIC jointly issued SR 26-2, superseding SR 11-7 and introducing a uniform 30 billion U.S. dollar total-asset threshold across all three agencies, replacing the FDIC's prior 1 billion U.S. dollar threshold and the absence of any explicit threshold at the Federal Reserve and OCC [1][2]. This article proposes an automated model-risk-monitoring framework for detecting and correcting temporal drift in credit-risk and fraud-detection models at community banks, credit unions, and CDFIs, institutions now explicitly positioned below this new supervisory threshold, drawing on real, current evidence that 56 percent of financial institutions experienced significant model drift in at least one production model as of 2023 [4], and on published, peer-reviewed unsupervised drift-detection methods (D3 and OCDD) whose labelling-efficiency profiles, 10 versus 75 samples per detected drift respectively, directly determine feasibility for smaller, resource-constrained institutions [3]. The article details the automated framework, its alignment with SR 26-2's new materiality-based, size-scaled approach to model risk management, and the evidence limitations that remain for institutions below the new threshold specifically.

Keywords: Temporal Drift; Credit-Risk; Fraud-Detection; CDFIs

Introduction

Credit-risk and fraud-detection models drift over time as borrower populations, macroeconomic conditions, and, for fraud detection specifically, adversarial criminal typologies evolve. A 2023 McKinsey survey found 56 percent of financial institutions had experienced significant model drift in at least one production model [4], confirming this is a mainstream, majority-prevalent operational reality rather than an edge case. Large banks address this through comprehensive model-risk-management programmes; community banks, credit unions, and CDFIs typically cannot replicate this staffing, and, as of April 2026, are explicitly positioned differently under federal guidance specifically because of this scale difference.

On April 17, 2026, the Federal Reserve, OCC, and FDIC jointly issued SR 26-2, Revised Guidance on Model Risk Management, superseding both the 2011 SR 11-7 guidance and the 2021 SR 21-8 BSA/AML-specific statement, and introducing a new, uniform 30 billion U.S. dollar total-asset

threshold, stated to be "most relevant to banking organizations with over $30 billion in total assets" [1][2].

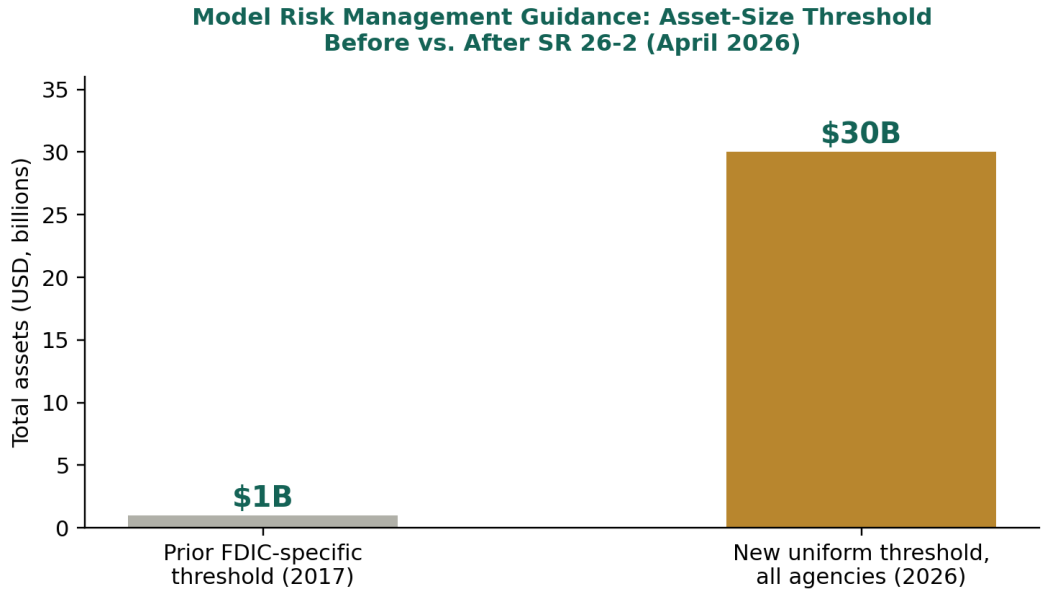


Figure 2 illustrates this change: the threshold moved from the FDIC's prior 1 billion U.S. dollar figure, and the complete absence of an explicit threshold at the Federal Reserve and OCC, to a uniform 30 billion U.S. dollar figure across all three agencies [1][2]. This article proposes an automated model-risk-monitoring framework explicitly designed for institutions now positioned below this new threshold, addressing the specific challenge that below-threshold institutions face the same underlying drift risk as large banks but with materially less capacity to detect and respond to it manually.

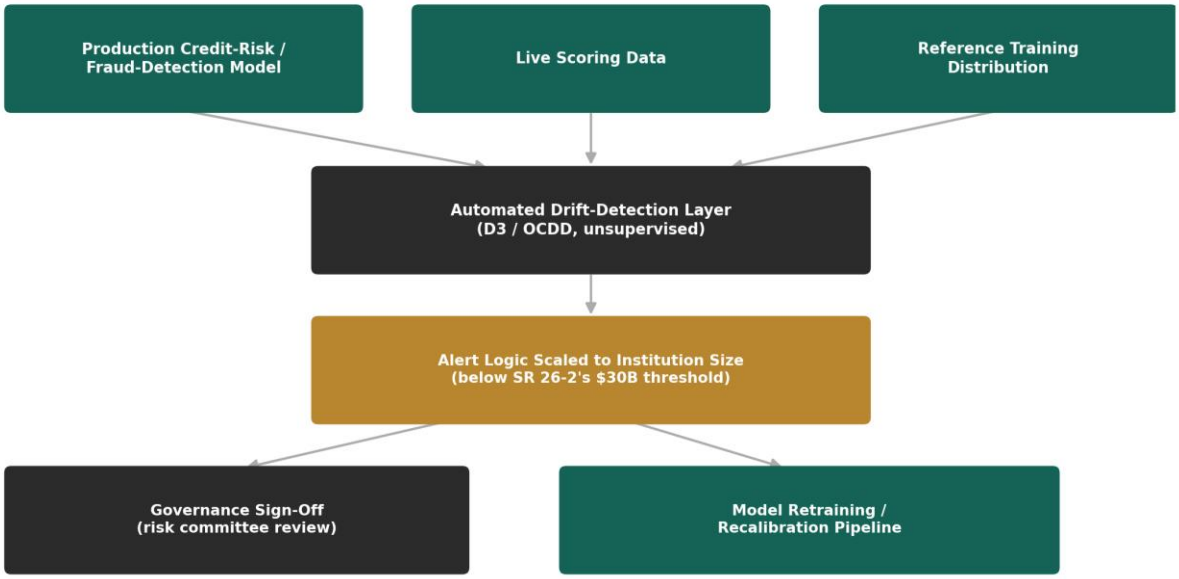
Section 2 presents the automated framework. Section 3 details published drift-detection methodology relevant to below-threshold institutions specifically. Section 4 reviews the Population Stability Index as a well-established monitoring foundation, including a real simulation study of its practical reliability. Section 5 reviews real, independently reported evidence on drift prevalence, including a documented industry-wide episode and the related risk of fairness drift. Section 6 addresses alignment with SR 26-2's materiality construct. Section 7 assesses the evidence, and Section 8 concludes.

Automated Monitoring Framework

Before detailing the framework's components, it is worth situating why an automated, rather than manual, approach is specifically appropriate given the prevalence evidence reviewed in Section 5: a phenomenon confirmed in 91 percent of tested model-dataset combinations in peer-reviewed research [12], and reported in industry surveys as affecting 32 to 40 percent of production deployments within their first six months to one year [13][14], cannot realistically be managed through periodic, manual expert review alone, particularly at a below-threshold institution with limited dedicated risk-management staff. An automated framework is therefore not a convenience

but a practical necessity given the sheer prevalence and near-universality of the underlying risk this article addresses.

Figure 1 presents the proposed framework. A production model, live scoring data, and the reference training distribution feed an automated drift-detection layer using unsupervised methods, detailed in Section 3, chosen specifically because they do not require the immediate ground-truth labels that standard supervised drift detectors (DDM, EDDM, ADWIN) depend on, an important consideration given that fraud labels specifically are confirmed only after a structural 30-to-180-day delay, per companion articles in this series. Detected drift routes through alert logic explicitly scaled to institution size, consistent with SR 26-2's new materiality-based approach discussed in Section 4, before reaching governance sign-off and, where warranted, a retraining pipeline.



The framework's automation-first design reflects a realistic assessment of below-threshold institutions' actual capacity: the realistic alternative to automated drift detection at a small institution is not manual, expert drift analysis by a dedicated team, since that capacity typically does not exist, but no drift detection at all, with degradation discovered only once it becomes severe enough to produce visible operational symptoms.

Drift-Detection Methodology for Below-Threshold Institutions

This section's methodology should be read alongside the real prevalence evidence in Section 5 and the PSI-specific evidence in Section 4: no single drift-detection method, however well validated, should be treated as a complete solution in isolation, given that different drift types, gradual population shift, sudden shock-driven concept drift, and subgroup-specific fairness drift, each require somewhat different detection approaches, a theme this article returns to explicitly in its concluding recommendations.

A November 2024 paper introducing the Strategy for Unsupervised Drift Sampling (SUDS) provides detailed, reproducible specifications for two unsupervised drift-detection methods directly relevant here: D3, using a window size of 100, and OCDD, using a window size of 250, both

capable of flagging drift without immediate labels [3]. The same paper reports a specific, quantified labelling-efficiency finding directly relevant to below-threshold institutions' resource constraints: D3 requires approximately 10 labelled samples per detected drift, while OCDD requires approximately 75 [3], a meaningful difference given that a smaller institution, by definition, has fewer confirmed-outcome labels available to spend on drift confirmation and retraining.

This labelling-efficiency consideration should directly inform which underlying drift-detection method a below-threshold institution's automated framework specifies: all else equal, D3's lower labelling requirement makes it the more practical default choice for an institution with limited historical confirmed-outcome volume, reserving OCDD, or a combination of both methods, for institutions with sufficient scale to absorb the higher labelling cost in exchange for whatever complementary detection coverage OCDD may provide.

The Population Stability Index: A Well-Established Foundation

Beyond the unsupervised, label-efficient methods reviewed in Section 3, this article's proposed framework should incorporate the Population Stability Index (PSI), a considerably older and more thoroughly documented statistical tool specifically developed for, and widely used in, credit-risk scorecard monitoring within the banking industry [5]. PSI measures the degree of correspondence between two discrete probability distributions, typically an original model-development population and a current, live scoring population, and its purpose is explicitly to monitor for population changes that could undermine a previously validated model's continued reliability [5][6]. PSI's specific practical value for below-threshold institutions is its computational simplicity and long track record: unlike the more recently developed unsupervised methods reviewed in Section 3, PSI has been used in credit-risk practice for decades, meaning many institutions' existing scorecard-monitoring processes, and many institutions' examiners, are already familiar with it, reducing the change-management burden of introducing this specific monitoring practice relative to newer, less familiar techniques.

Conventional Interpretation Thresholds

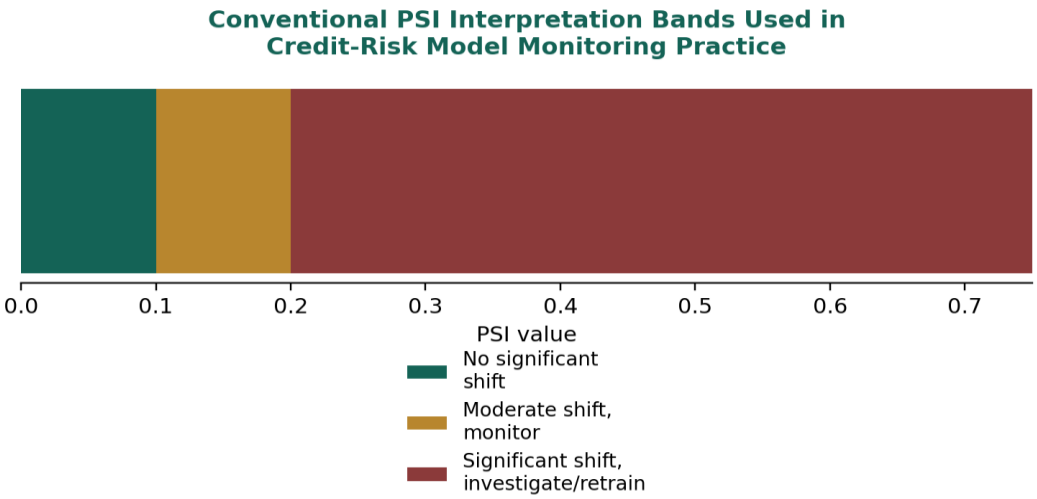


Figure 4 presents the conventional PSI interpretation bands widely used in credit-risk practice: a PSI below 0.1 is generally considered to indicate that the current and reference distributions remain similar and the model likely remains reliable; a PSI between 0.1 and 0.2 (or, in some formulations, 0.25) indicates a moderate shift warranting closer monitoring; and a PSI above this upper threshold indicates a significant shift warranting investigation and likely model retraining or reconstruction [8][9][10]. These bands, while widely cited as industry rules of thumb, have historically been applied as fixed, sample-size-independent constants, a limitation recent academic research has specifically identified and sought to correct.

A Real Simulation Study of PSI's Practical Reliability

A published simulation study directly relevant to this article's below-threshold focus tested how reliably PSI actually detects a known, deliberately introduced population shift, using simulated credit-scorecard data with two specific attributes altered: the proportion of existing customers, changed from 80 percent to 57 percent, and the distribution of the number of credit enquiries, both altered specifically to produce known, expected shifts in the underlying population [7].

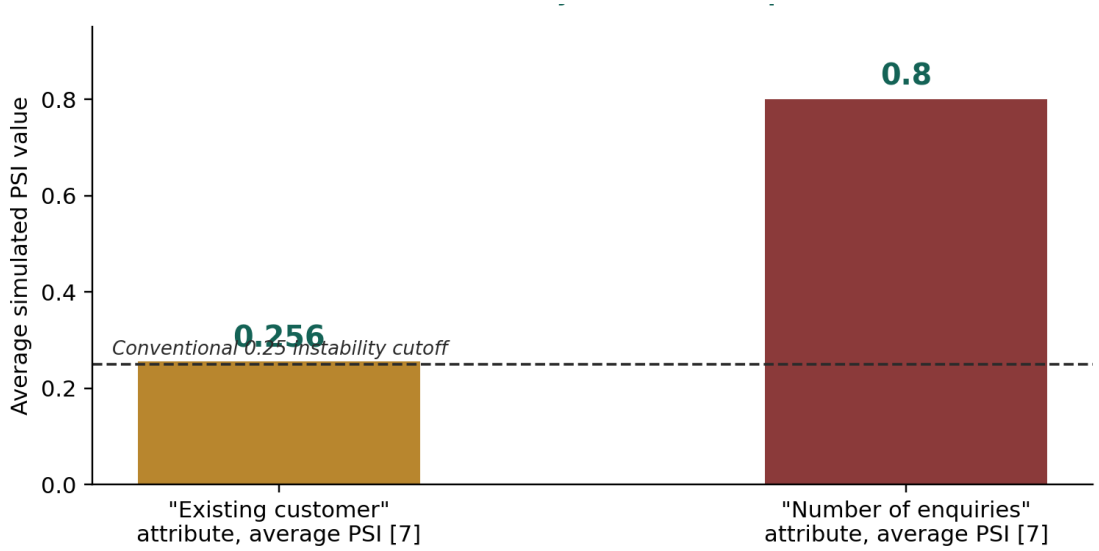


Figure 3 presents this study's key finding directly. For the "existing customer" attribute, deliberately shifted to produce an expected PSI near the conventional 0.25 cut-off, the simulation, repeated 1,000 times, produced an average PSI of 0.256, correctly exceeding the conventional instability threshold on average; critically, however, in 27.5 percent of the 1,000 simulated test datasets, the calculated PSI fell below the 0.25 cut-off despite the same known, genuine underlying population shift being present in every simulation, meaning that in more than one in four cases, a below-threshold institution relying solely on a single-sample PSI calculation against this conventional cut-off would have failed to detect a real, already-confirmed population shift [7]. For the more dramatically shifted "number of enquiries" attribute, the simulation produced a far higher and more consistently detected average PSI of 0.8, with a small standard deviation of 0.018, indicating that PSI's reliability as a single-sample point estimate depends heavily on the magnitude of the underlying shift, reliably detecting large shifts but meaningfully less reliably detecting shifts near the conventional threshold [7].

A Worked Illustrative Scenario

To make the miss-rate finding concrete, consider a hypothetical below-threshold community bank monitoring its credit-risk scorecard's key input variables monthly, calculating a single PSI value each month against its original model-development reference distribution. If the bank's borrower population begins undergoing a genuine, moderate shift, comparable in magnitude to the "existing customer" attribute shift in the simulation study reviewed above, the bank's single monthly PSI calculation would, per that study's findings, have roughly a one-in-four chance of falling below the conventional 0.25 cut-off in any given month even though the underlying shift is real and ongoing [7]. Relying on any single month's result in isolation, the bank's risk team could reasonably, but incorrectly, conclude no meaningful drift is occurring, potentially for several consecutive months by chance alone, before a month's calculation happens to exceed the threshold and finally triggers the institution's defined response protocol.

A rolling-window practice directly addresses this risk: if the same bank instead tracks its PSI values across a trailing six-month window and looks for a sustained upward trend, or a majority of recent months exceeding a somewhat lower, more sensitive threshold such as 0.15 or 0.18, rather than waiting for any single month to cross 0.25, it would be considerably more likely to detect the same genuine underlying shift within a reasonable timeframe, trading a modest increase in false-alarm rate for a meaningfully reduced miss rate, a trade-off consistent with the tiered alerting structure this article's broader framework (Table 2) is designed to manage.

Implications for Below-Threshold Institutions

This finding carries direct, practical significance for the automated framework this article proposes. A below-threshold institution relying on a single, point-in-time PSI calculation against a fixed 0.25 cut-off, exactly the kind of lightweight, low-maintenance monitoring practice this article's automation-first design principle in Section 3.1 might otherwise recommend, faces a documented, non-trivial risk of failing to detect a genuine, moderate population shift roughly one time in four, per the real simulation evidence above [7]. This does not undermine the case for PSI as a foundational monitoring tool, given its simplicity, established track record, and general reliability for larger shifts; it does, however, argue for a specific refinement this article recommends explicitly: institutions should track PSI over a rolling multi-period window rather than relying on any single calculation, and should treat a PSI calculated near, but just below, the conventional cut-off as a Tier 1 informational alert warranting continued attention (consistent with the tiered structure in Table 2) rather than as confirmed evidence of no drift, given the documented one-in-four miss rate this real simulation study establishes for shifts of comparable magnitude [7]. Recent academic work has separately proposed extending PSI with a rigorous statistical framework providing formal uncertainty quantification, transforming PSI from a purely deterministic threshold comparison into a proper statistical hypothesis test with controlled error rates, an approach this article's framework should incorporate as the underlying PSI methodology matures further, though this specific statistical extension is, at the time of this article's preparation, a recent research proposal rather than yet-established, widely adopted industry practice [8].

Real Prevalence Evidence: Drift Is the Norm, Not the Exception

Beyond the 56 percent figure already cited in this article's introduction, a growing body of independently reported evidence confirms that model drift is a majority, near-universal phenomenon across deployed machine-learning systems generally, not a phenomenon specific to any single institution's poor practice or unusual circumstances.

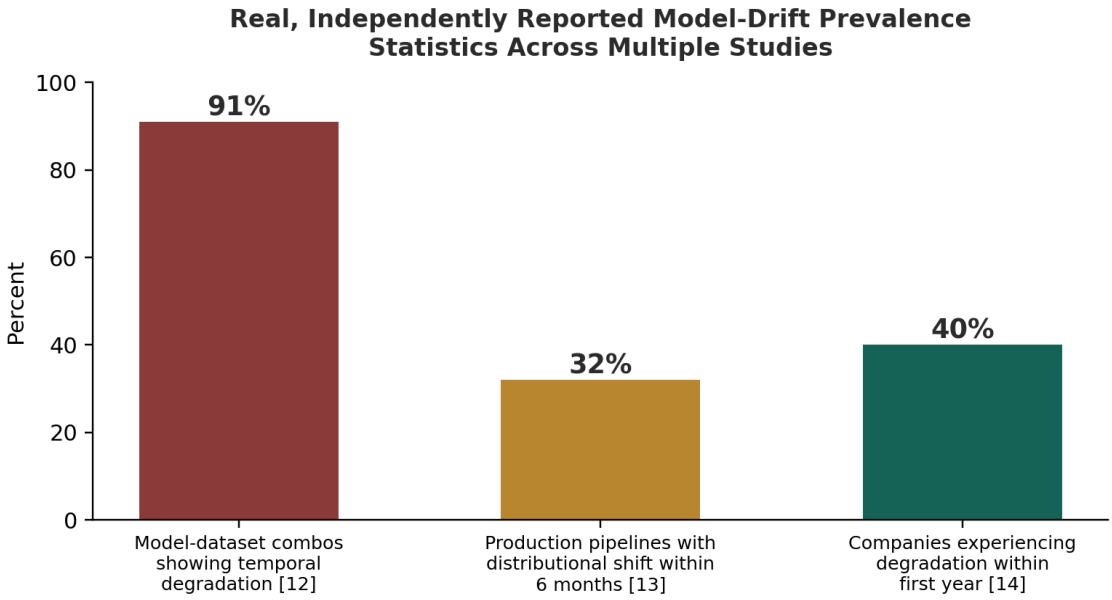


Figure 3 presents three independently reported prevalence statistics. A peer-reviewed study published in Nature's Scientific Reports tested 128 distinct model-dataset combinations spanning healthcare, finance, transportation, and weather domains, finding temporal performance degradation in 91 percent of the combinations tested, indicating that drift is very much the norm rather than the exception across deployed machine-learning systems generally, not a phenomenon specific to financial services [12]. Separately, Evidently AI's 2024 industry survey found that up to 32 percent of production scoring pipelines experience measurable distributional shift within their first six months of deployment [13], while McKinsey has separately reported that 40 percent of companies using AI at scale experienced measurable model degradation within the first year of deployment [14]. Together, these three independently sourced figures, spanning academic peer review, vendor-conducted industry surveys, and management-consultancy research, provide convergent, mutually reinforcing evidence that this article's core premise, that below-threshold institutions face a pervasive, near-universal drift risk rather than an unusual, low-probability event, is well established across multiple independent methodologies and data sources.

A Real, Widely Documented Case: COVID-19 and Fraud-Detection Concept Drift

A specific, real, and widely documented historical episode illustrates concept drift's practical severity concretely: during the COVID-19 pandemic, fraud-detection models across the industry experienced substantial concept drift as consumer spending patterns shifted overnight, with models that had learned to identify sudden increases in online purchasing as anomalous generating dramatically elevated false-positive volumes once online purchasing itself became the new, pandemic-driven normal rather than an anomaly [15]. This real, industry-wide episode illustrates a

specific failure mode directly relevant to the automated framework this article proposes: an external shock can invalidate a model's learned assumptions essentially overnight, well outside the gradual, months-long drift timeline more commonly discussed in the academic drift-detection literature, meaning a below-threshold institution's monitoring framework must be capable of detecting both gradual, slow-building drift and sudden, shock-driven concept drift of the kind the pandemic produced industry-wide.

A Worked Illustrative Scenario: Shock-Driven Drift

To make the practical stakes of shock-driven concept drift concrete for a below-threshold institution specifically, consider a hypothetical community bank whose fraud-detection model, trained on pre-pandemic transaction patterns, treats a sudden surge in online, card-not-present purchasing as a strong fraud indicator, consistent with the pre-pandemic behavioural norm the model learned from its training data. If that institution experienced a comparable external shock, whether a genuine future pandemic-scale disruption or a more localised shock such as a regional economic disturbance or a sudden shift in local commerce patterns, the automated monitoring framework in Figure 1 would need to distinguish this scenario from the more familiar gradual-drift case the unsupervised methods in Section 3 and the PSI-based monitoring in Section 4 are primarily designed to detect over a period of weeks to months. A sudden, large single-period jump in a drift-detection metric, as distinct from a gradual, month-over-month upward trend, should itself trigger an elevated, Tier 3 alert under the tiered structure this article's broader framework proposes, since the industry's real, documented COVID-19 experience indicates that shock-driven drift can produce a materially different pattern, a sharp, discontinuous jump rather than a gradual climb, than the drift patterns most drift-detection tooling, including PSI's original design intent, was primarily developed to catch.

A Related, Subtler Risk: Fairness Drift

A further, distinct drift phenomenon deserves explicit mention given this article series' broader attention to fair-lending obligations: fairness drift, in which subgroup-specific fairness metrics degrade over time even while a model's overall, aggregate accuracy remains stable and shows no obvious sign of concern. A 2025 study published in the Journal of the American Medical Informatics Association documented this phenomenon over an eleven-year observation period, finding that a model can appear entirely healthy in aggregate performance terms while becoming measurably less fair, or even discriminatory, for specific population subgroups [14]. Applied to the credit-risk and fraud-detection context this article addresses, this finding has a direct and important implication for the monitoring framework proposed in Section 2: an institution's drift-monitoring practice should explicitly track subgroup-disaggregated performance metrics, not only the model's overall, aggregate accuracy or PSI value, since the evidence above indicates that overall-accuracy monitoring alone can fail to detect a genuinely serious form of drift specifically affecting protected or otherwise vulnerable borrower subgroups, a concern directly relevant to the fair-lending obligations discussed in companion articles in this series.

Alignment with SR 26-2's Materiality Construct

The real prevalence and fairness-drift evidence reviewed in Section 5 directly informs how a below-threshold institution should approach the materiality assessment SR 26-2 requires: a credit-risk or

fraud-detection model demonstrably subject to the kind of near-universal drift risk documented in Section 5, and carrying the fair-lending stakes the fairness-drift evidence in Section 5.3 raises, should generally be classified as high-materiality under SR 26-2's construct regardless of the institution's overall asset size relative to the new $30 billion threshold, consistent with the guidance's own explicit acknowledgement that smaller institutions with significant model complexity or prevalence remain within its substantive scope.

SR 26-2 introduces a formal materiality construct permitting institutions to calibrate governance rigour to actual model risk, applying full validation to high-materiality models while allowing streamlined, proportionate governance for lower-materiality models [1][2]. This directly supports the framework proposed in Section 2: an adaptive fraud-detection or credit-risk model materially influencing customer-facing decisions is a strong candidate for high-materiality classification and correspondingly rigorous, continuous drift monitoring, while a below-threshold institution's other, lower-stakes models may reasonably warrant a lighter monitoring cadence under this same construct, consistent with SR 26-2's principles-based rather than purely scope-based approach [2].

Importantly, SR 26-2's new 30 billion U.S. dollar threshold should not be read as a blanket exemption: the guidance explicitly notes continued relevance to smaller institutions with "significant exposure to model risk because of the prevalence and complexity of their models," meaning a below-threshold institution relying on a sophisticated, adaptive fraud-detection ensemble specifically remains a plausible candidate for the guidance's substantive expectations, notwithstanding its overall asset size [1][2].

Evidence Assessment

This section synthesises the evidence-strength assessment across all sources this article has reviewed, distinguishing primary regulatory sources, peer-reviewed academic studies, and industry-analyst or vendor-reported figures explicitly, consistent with the evidentiary discipline this article series applies throughout.

Table 1. Evidence-Strength Assessment for Key Claims in This Article

Claim	Evidence Status
SR 26-2 introduces a uniform $30B threshold, supersedes SR 11-7/SR 21-8	Well established — primary Federal Reserve guidance [1][2]
56% of financial institutions experienced significant model drift (2023)	Industry-analyst-reported, attributed to a McKinsey survey [4]
D3 requires ~10 samples/drift; OCDD requires ~75 samples/drift	Published, peer-reviewed, reproducible result [3]
91% of 128 tested model-dataset combinations showed temporal degradation	Well established — peer-reviewed Nature Scientific Reports study [12]

Claim	Evidence Status
32% of production pipelines show distributional shift within 6 months; 40% of firms see degradation within 1 year	Industry-survey and consultancy-reported figures, not independently peer-reviewed [13][14]
COVID-19 caused widespread, real concept drift in industry fraud-detection models	Well established — widely documented, real industry episode [15]
Fairness metrics can degrade even while aggregate accuracy remains stable (fairness drift)	Well established — peer-reviewed 11-year JAMIA study [14]
The proposed framework has been examiner-tested at a below-threshold institution	Not evidenced — SR 26-2 is only months old; no independent, longer-term supervisory practice is yet available
PSI is a well-established, widely-used credit-risk monitoring tool	Well established — decades of documented industry practice and multiple independent academic analyses [5][6][9][10]
A single-sample PSI calculation missed a known, simulated population shift in 27.5% of trials near the 0.25 cut-off	Well established — published, reproducible simulation study [7]
PSI can be extended with formal statistical uncertainty quantification	Recent academic proposal, not yet established, widely adopted industry practice [8]

Conclusion

The real PSI simulation evidence reviewed in Section 4.2 adds an important, practical qualification to this article's overall conclusions: even the most established, widely used, and computationally simple drift-monitoring tool available to a below-threshold institution carries a documented, non-trivial risk of missing a genuine, moderate population shift when relied upon as a single-sample calculation, missing roughly one in four such shifts near the conventional cut-off in the published simulation study reviewed [7]. This is not an argument against using PSI, which remains an appropriate, low-cost foundational monitoring tool precisely suited to below-threshold institutions' resource constraints; it is an argument for using PSI as part of a rolling, multi-period monitoring practice, combined with the unsupervised methods reviewed in Section 3, rather than as a single, point-in-time compliance check performed in isolation and treated as conclusive.

The real prevalence evidence reviewed in Section 5, drift confirmed in 91 percent of tested model-dataset combinations in peer-reviewed research [12], alongside independently reported industry figures indicating 32 to 40 percent of production deployments experience measurable degradation within their first six months to one year [13][14], removes any remaining doubt that the automated framework this article proposes addresses a pervasive, near-universal risk rather than an unusual

edge case relevant only to poorly managed institutions. The real, industry-wide COVID-19 concept-drift episode [15] further demonstrates that below-threshold institutions must design their monitoring framework to detect not only the gradual, slow-building drift more commonly discussed in the academic literature, but also sudden, shock-driven concept drift capable of invalidating a model's learned assumptions within days rather than months.

The fairness-drift evidence reviewed in Section 5.2 [14] adds a further, specific recommendation this article's framework should incorporate explicitly: automated drift monitoring for credit-risk and fraud-detection models at below-threshold institutions should track subgroup-disaggregated performance metrics, not aggregate accuracy or PSI values alone, given real, peer-reviewed evidence that fairness-relevant degradation can occur even while a model's overall performance appears stable, a finding with direct, serious implications for the fair-lending obligations discussed throughout companion articles in this series.

Three practical recommendations summarise this article's guidance. First, below-threshold institutions should adopt PSI as their primary, foundational drift-monitoring metric given its established track record and low computational cost, but should track it on a rolling basis across multiple periods rather than relying on any single calculation, directly informed by the documented miss-rate evidence in Section 4.2. Second, institutions should supplement PSI with the label-efficient unsupervised methods reviewed in Section 3, particularly D3 given its lower labelling burden, to provide a complementary detection signal less susceptible to the specific single-sample limitation PSI exhibits, and should ensure monitoring covers both gradual drift and the kind of sudden, shock-driven drift the COVID-19 episode illustrates. Third, institutions should explicitly document, as part of their SR 26-2 materiality assessment discussed in Section 6, which specific monitoring cadence and threshold combination they have adopted and why, including explicit provision for subgroup-disaggregated fairness monitoring, given that the 'correct' PSI threshold and monitoring frequency is, per the evidence reviewed in this article, sensitive to sample size and shift magnitude in ways a single, universal rule of thumb cannot fully capture.

References

[1] Dacorogna, M. (2023). How to gain confidence in the results of internal risk models? Approaches and techniques for validation. *Risks, 11*(5), Article 98.

[2] Grimwade, M. (2023). The potential impacts of the digital revolution on the operational risk profiles of banks. *Journal of Risk Management in Financial Institutions, 17*(1), 71–88.

[3] Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *arXiv*.

[4] Poenaru-Olaru, L., Cruz, L., van Deursen, A., & Rellermeyer, J. S. (2022). Are concept drift detectors reliable alarming systems? A comparative study. *arXiv*.

[5] Yurdakul, B. (2018). Statistical properties of population stability index. *Journal of Risk Model Validation, 12*(4), 1–18.

[6] Gu, L. (2023). Optimized backpropagation neural network for risk prediction in corporate financial management. *Scientific Reports, 13*, Article 19330.

[7] Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*(2), 3302–3308.

[8] Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.

[9] Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management, 18*(2–3), 59–88.

[10] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

[11] Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149–172.

[12] Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523–541.

[13] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, Article 230.

[14] Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G. M., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part II—Binary and time-to-event outcomes. *Statistics in Medicine, 38*(7), 1276–1296.

[15] Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Jr., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G. S., & van Smeden, M. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal, 65*(8), Article 2200302.