

Failing FDA drug inspections are recorded later and less often than passing ones, biasing models trained and monitored on them

Phani Kumar Balagam

balagam.phani@gmail.com

Independent Researcher

Article

Keywords: FDA inspections, label availability, informative censoring, model monitoring, regulatory data quality, machine learning

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11122103/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Inspection classifications published by the US Food and Drug Administration are an attractive label for machine learning: public, dated, keyed to facilities, and running to hundreds of thousands of records. We asked whether the label's availability depends on its value, and what that costs anyone who models or monitors with it. Among 10,201 drug inspections that ended between 2018 and 2023 with at least three years to close, 74.1% of passing inspections carry a recorded final classification decision against 26.2% of failing ones, a relative risk of 2.83 (95% CI 2.54–3.21); among those that close, failures take 3.86 times longer (median 332 against 86 days; 95% CI 2.94–4.93). Rebuilding an inspection-outcome model so that a prior outcome counts only once its decision is recorded costs 0.047 to 0.081 in area under the receiver operating characteristic curve, with non-overlapping intervals. For monitoring a deployed model, the naive lagged observed-to-expected ratio is 5.7% of its full-information value and inverse-probability-of-censoring weighting 6.3%; weighting's identifying assumption fails by 0.499, and landmarking at two years recovers 27.6%. Constructing features on the availability clock repairs the first consequence; nothing we tested repairs the second. A single published posting-date field would resolve both.

Introduction

FDA publishes the outcome of every inspection it conducts. The Data Dashboard exposes them through an API keyed on facility, dated to the day, and classified on a three-level scale: NAI where no objectionable conditions were found, VAI where they were but no action follows, and OAI where regulatory or administrative action is recommended [1]. For anyone building models of regulatory risk this is a rare thing, a large labeled corpus about corporate behavior that needs no access agreement. A small literature already builds on FDA outcome data of this kind, including quantitative analyses of recalls [2, 3], of good clinical practice inspection classifications [4], and of warning letters [5, 6].

The label has a property that is not documented and, as far as we can find, not previously measured: **whether it is available depends on what it is**. An inspection's classification enters the record when FDA logs the FMD-145 final classification decision. Passing inspections are logged promptly and usually. Failing inspections are logged slowly and often not at all. A model or a monitor that consumes this feed therefore sees a version of the world in which failures are rare, and sees it consistently enough that nothing in the data announces the problem.

This is not a new statistical phenomenon. It is informative censoring in survival analysis [7, 8] and verification bias in diagnostic-accuracy evaluation [9], and in its outcome-dependent form it is the case Zhou [10] separates from the tractable one. Tsiatis [11] proved fifty years ago that a dependent-censoring model cannot be distinguished from an independent one by observed data alone. What is new here is the measurement, in a widely used public regulatory dataset, and the demonstration that two things people routinely do with that dataset are damaged by it in ways that are large, quantifiable, and in one case not repairable.

The contributions are three. **The measurement** (*Outcome-dependent availability*): outcome-dependent availability in FDA drug-inspection classifications, on a cohort chosen so that right-censoring and historical under-recording do not confound it. **Leakage in supervised models** (*Availability leakage in supervised models*): features built from prior classifications are unavailable at prediction time, and enforcing availability costs accuracy and removes the facility's compliance history as a usable signal. **Monitoring is not repairable by the standard corrections** (*Monitoring is not repairable by the standard corrections*): the identifying assumption for inverse-probability-of-censoring weighting fails by a measured margin, and landmarking stops well short of the truth because the missing failures are a mass that never arrives rather than a delay.

Related work

The phenomenon has names in three literatures and no name in the one that most often uses this data. In survival analysis it is informative or dependent censoring [7, 8]; when it is outcome-dependent it is non-ignorable, the regime in which Scharfstein and colleagues [12] recommend a sensitivity parameter treated as known rather than an estimator. In diagnostic-accuracy evaluation it is verification bias [9], with the outcome-dependent case treated separately [10, 13]. Tsiatis [11] is the reason none of these is a repair: the dependent and independent models are not distinguishable from observed data.

In stream learning the delay itself is well studied as verification latency [14, 15, 16], and delayed-label evaluation has a reference treatment in Grzenda and colleagues [17]. That literature generally treats the delay as a property of the instance rather than of its label: Grzenda and colleagues [17] model a per-instance latency without dependence on the true class, and Dal Pozzolo and colleagues [18] assume a constant verification latency. Souza and colleagues [19] make the adjacent point that benchmark conclusions are protocol dependent. The closest structural precedent is Chapelle [20], where a conversion is confirmed after a random delay and a non-conversion is never confirmed at all; that asymmetry is the same shape as the one measured here.

On monitoring deployed models, Davis and colleagues [21] established calibration drift as the failure mode to watch and Kore and colleagues [22] showed on real imaging data that performance monitoring is a poor proxy for drift. Both assume the labels arrive. On FDA outcome data specifically, the analyses cited above model recalls [2, 3], describe good clinical practice inspection classifications [4], and content-analyze warning letters [5, 6]. We are not aware of prior work measuring the availability process of these labels.

Results

Outcome-dependent availability

Table 1 gives the measurement. Both mechanisms are large and both point the same way. A passing inspection is 2.83 times more likely to have its outcome recorded at all, and when a failing inspection is

recorded it takes almost four times as long.

Table 1
Outcome-dependent availability, primary cohort: inspections ending 1 January 2018 to 8 September 2023 with at least three years of follow-up. Confidence intervals by bootstrap, 2,000 resamples.

n	not OAI	OAI	ratio (95% CI)
	9,384	817	
ever recorded	74.1%	26.2%	2.83 (2.54–3.21)
median days to record, among recorded	86	332	3.86 (2.94–4.93)

The consequence for anything reading this feed is direct. At a 90-day horizon the Kaplan-Meier availability is 22.4% for passing inspections and 1.5% for failing ones (Fig. 1, left). A monitor consulting three-month-old labels is looking at a sample in which failures are under-represented by roughly an order of magnitude.

The effect is not uniform across FDA's programs. Within Drug Quality Assurance, 82.4% of passing inspections are recorded against 24.2% of failing ones; within Bioresearch Monitoring the same figures are 63.3% and 45.3%. The gap is widest in exactly the program most likely to be modeled.

We do not know FDA's reason and do not speculate. A plausible mechanism, consistent with the direction and with the OAI definition, is that an adverse classification initiates follow-up whose conclusion is what the FMD-145 record captures, so the record waits for the enforcement process rather than for the inspection. Nothing in the published data distinguishes that from alternatives, and the analysis below does not depend on which is true.

Availability leakage in supervised models

A classification code does not exist until the decision is recorded, which for an OAI is a median of 332 days after the inspection ends and, more often than not, never. At scoring time T a prior OAI from T minus 200 days is not an OAI in anything a deployed system could query.

The clock removes most of the history. Only 31.6% of prior inspections in the window have a usable outcome, and 75.3% of rows have no usable prior outcome at all, against 100% assumed by the unlocked construction. On the 2013–2016 training window, deployment AUROC falls from 0.7696 to 0.6882, a gap of 0.0814 with non-overlapping intervals (Fig. 2).

That estimate mixes two causes, because before 2018 the export is materially incomplete: 72.5% of pre-2018 records carry no recorded decision against 27.9% from 2018 onward, and the training window for those figures is 2013–2016. Retraining on 2018–2019, where recording is mature, the gap is 0.0466 (0.7880 unlocked against 0.7414 clocked), again with non-overlapping intervals. We take 0.047 as the

conservative figure, applying to a model trained where recording is mature, and 0.081 as the figure applying to a model trained on an era whose export is materially incomplete - which is the case for anyone training on the series as published back to 2008.

Both figures come from a strict clock: a prior outcome counts only where the export carries a recorded decision dated before the prediction date. The permissive alternative is to impute a closure date for records that carry none, drawing from the outcome-conditional latency distribution, and under it the deployment gap falls to 0.0116. That contrast is the largest sensitivity in this section, and we report the strict arm as primary for two reasons. Roughly three quarters of failing inspections never record a decision at any horizon (*Outcome-dependent availability*), so imputing one converts a permanent absence into a finite delay and assumes away the mechanism the rest of the paper measures. And a deployed system faces the strict clock whatever the truth of the matter, because it can query only what the export actually carries.

What changes is not only the number. Under the clock the largest coefficients are the project area FDA assigned to the inspection (-1.07 and - 0.53), whether the facility has no recorded prior history at all (0.54), and the facility's country (0.35). No outcome-derived feature appears in the top five. The apparent predictive value of a facility's compliance record, in this data source, rests substantially on knowing outcomes that were not knowable.

This consequence is fixable, and cheaply. Build history features on the availability clock: admit a prior record's outcome only if the source carries a decision date earlier than the prediction date. The construction is in the supplied analysis code. The cost is a smaller and more honest model.

Monitoring is not repairable by the standard corrections

Across 88 monthly evaluation windows the naive lagged ratio averages 0.057 of its full-information value, computed on the same window with all labels. The monitored quantity is not a noisy version of the model's calibration; it is a different quantity roughly a factor of seventeen below it.

Inverse probability of censoring weighting is the standard correction for dependent censoring [23]. It identifies the target only under conditionally independent censoring; Willems and colleagues [24] state the requirement as the hazard of censoring at time t being independent of the event time given the covariate history, that is, that there should be no unmeasured confounders. Fitting a model for whether a label arrives within 365 days, on the covariates a monitor can see, gives AUC 0.708; adding the true outcome raises it to 0.769. After adjusting for every covariate, the mean residual arrival probability is plus 0.039 for passing inspections and minus 0.461 for failing ones, a residual gap of 0.499. The assumption does not approximately hold; it fails by half the range of a probability. Applied anyway, weighting does not help: 0.063 of the full-information value against 0.057 naive (Table 2, Fig. 3).

Table 2

Estimators of the observed-to-expected ratio, each compared against the full-information value on the same window. Mean over 88 monthly windows for the naive and weighted estimators and 89 for the landmark estimators, January 2019 to May 2026. A value of 1.0 would be unbiased.

estimator	mean ratio to full information	SD
naive (labels arrived so far)	0.057	0.050
inverse probability of censoring weighting	0.063	0.061
landmarking, H = 180 d	0.161	0.088
landmarking, H = 365 d	0.231	0.110
landmarking, H = 730 d	0.276	0.118

Landmarking helps and then stops. It moves the estimate from 0.057 to 0.276 at H = 730 days, and cannot go much further. A longer wait recovers a delayed label but not an absent one, and roughly three quarters of failing inspections have no recorded decision at any horizon. That mass does not shrink with H.

Discussion

A numeric acceptance criterion on a label-dependent metric cannot be met in this system. The European Medicines Agency's reflection paper on artificial intelligence in the medicinal product lifecycle asks that thresholds for acceptable model performance be clearly defined [25]. A threshold on the observed-to-expected ratio computed from arriving labels is a threshold on a quantity whose level is set by the label pipeline, not by the model, and no estimator we tested recovers the level: the naive estimate sits at 5.7% of the truth and two years of landmarking reaches 27.6%. Monitoring for *change* may still be possible, because a stable bias shifts the level without erasing movement; monitoring for *conformance* is not, because the level is what a conformance criterion is about.

The practical guidance that follows is narrow and, we think, actionable. First, build history features on the availability clock and report the discrimination that survives, rather than the discrimination obtained from outcomes the model could not have had. Second, do not express an acceptance criterion for a deployed model as an absolute threshold on a statistic computed from labels that arrive this way; express it, if at all, as a change criterion, and state the blind window explicitly. Third, when a monitoring plan depends on a label feed, document the feed's availability process with the same care as the model, because it is the part of the system that determines what the monitor can see.

FMD-145 is a proxy for availability, not availability. It is the recorded final classification decision, not the date the record appeared in the Dashboard, which is what a system consuming that API would face. The true availability latency is at least what we measure and plausibly longer. A published posting-date field would settle this, and would also make the whole problem tractable for everyone downstream.

Absence is not fully separable into its causes. We separate eras and report both, but within any era a missing decision may mean FDA has not finished, or that the export did not record it. For a model builder the distinction does not change what can be used; for interpreting FDA's process it matters, and we cannot resolve it from published data.

One agency, one product area. Drug inspections only. Whether device, food or biologics classifications behave the same way is checkable with the same method and not checked here.

The monitoring result is conditional on the model being monitored. The frozen model is one ordinary logistic regression on this substrate. The estimator comparison concerns the observed-to-expected ratio and would look similar for any model whose labels arrive this way, but we have measured it for one.

We offer no repair for the second problem. Availability-clocked features fix the leakage. Nothing we tested fixes the monitoring, and the non-identifiability result suggests nothing will without an external source of information about the missing labels.

FDA inspection classifications are a good label with a bad property. Among inspections ending 2018 to 2023 with three years to close, a failing inspection is 2.83 times less likely to have its outcome recorded than a passing one, and takes 3.86 times longer when it is. Models trained on prior classifications gain 0.047 to 0.081 AUROC from outcomes they could not have had. Monitors reading these labels see about a seventeenth of the calibration failure they are meant to catch, and the standard corrections do not recover it: the identifying assumption for inverse-probability-of-censoring weighting fails by a measured 0.499, and landmarking at two years reaches 27.6% because the missing failures never arrive at any horizon.

The first problem has a fix that costs accuracy and buys honesty. The second does not have one in the published data. Both would be substantially resolved by FDA publishing, alongside the classification, the date the classification became visible.

Methods

Data source and record keying

The FDA Data Dashboard supplied 341,046 inspection classification records and 280,130 inspection citation records, retrieved 7 September 2026 through the credentialed API [26]. Restricting to `ProductType = Drugs` gives 39,926 classification records: 18,053 NAI, 18,842 VAI and 3,031 OAI.

A record is keyed by (InspectionID, ProjectArea), not by InspectionID alone: 752 inspections carry more than one classification record because FDA classifies separately per project area, and 148 of those disagree on the outcome. Joining on InspectionID alone is many-to-many and inflates any analysis frame built from these tables.

The availability measure and the primary cohort

Each record may carry an FMD145Date, the recorded final classification decision. We treat an outcome as *available* on that date and unavailable otherwise, and define availability latency as the gap from InspectionEndDate. 85 records carry a decision date preceding the inspection and are dropped as data errors.

Two confounds are separated rather than pooled. The corpus was retrieved on 7 September 2026, so an inspection ending in 2026 has had months rather than years to close; cumulative availability is therefore estimated by Kaplan-Meier with administrative censoring at the retrieval date. Separately, the share of inspections carrying any recorded decision rises from 8.4% for 2008 inspections to 77.6% for 2019 inspections (Fig. 1, right). That is the export becoming more complete, not FDA taking fifteen years to close routine work. Pooling across the full corpus confounds the two and understates availability, so the primary cohort is restricted to inspections ending on or after 1 January 2018 with at least three years of follow-up: $n = 10201$, of which 817 are OAI and 9,384 are not. The full series is reported alongside it as a data-quality observation, not as the headline. Confidence intervals are by bootstrap with 2,000 resamples.

Prior analysis of this panel by the same author

An earlier manuscript by the present author — *Predicting the Inspector, Not the Plant: Regulatory Targeting and Explanation Instability in Pharmaceutical Inspection-Outcome Models*, submitted to *Therapeutic Innovation & Regulatory Science* on 13 September 2026 and under consideration at the time of writing, not published and not accepted for publication, and posted as a preprint at Research Square on 16 September 2026 [27] — builds a rolling-origin panel from the same two FDA Data Dashboard endpoints and reports this asymmetry inside its methods: a median of 314 days to a recorded closure for OAI against 91 days otherwise, and only 20.2% of OAI records carrying a closure date at all. That manuscript treats the asymmetry as a threat to validity and bounds its exposure — 212 rows, 0.54% of its panel, carry a prior OAI more recent than the 314-day median latency — rather than acting on it, on the stated ground that gating on closure date would remove most prior-OAI signal for reasons of data coverage rather than of science.

The present paper measures the mechanism instead of bounding it. It estimates the availability process itself on a cohort built for that purpose, quantifies what enforcing an availability clock costs a supervised model, and tests whether the standard corrections for outcome-dependent censoring repair a monitoring statistic computed through these labels. The two manuscripts share the data source and nothing else: no result, table, figure or claim is reported in both.

Quantifying availability leakage in a supervised model

Models built on this data commonly use a facility's prior inspection outcomes as features: how many prior OAI, how long since the last one, prior adverse rates. Those features are built from classification codes, and a classification code does not exist until the decision is recorded.

We quantify the consequence by rebuilding an inspection-outcome model twice under an otherwise identical protocol: one row per drug classification record, features confined to a trailing five-year window, L2 logistic regression fitted once and frozen, deployment 1 January 2020 to 31 May 2026 in every arm. The comparison is run on two training windows, because the completeness of the export changes what the measurement means: 1 October 2013 to 31 December 2016, an era in which the export is materially incomplete, and 1 January 2018 to 31 December 2019, an era in which recording is mature. The arms differ only in what the history features may see. The *unlocked* arm assumes every prior outcome is known, as is standard. The *availability-clocked* arm admits a prior outcome only if the export carries a recorded decision before the prediction date. Discrimination is compared by AUROC with 95% facility-level bootstrap intervals over 300 resamples, so that repeated inspections of one establishment are resampled together.

Estimators of the observed-to-expected ratio

The second use is monitoring a deployed model by comparing observed events to predicted ones. The observed-to-expected ratio, the mean outcome divided by the mean predicted risk over a trailing window, is the usual instrument, and calibration drift detected this way is a standard concern [21,28]. Here the numerator is computed from whichever labels have arrived, and those are the passing ones.

Four estimators are compared against the full-information value computed on the same window with all labels, over 88 monthly evaluation windows for the naive and weighted estimators and 89 for the landmark estimators, January 2019 to May 2026: the naive estimate from labels arrived so far; inverse probability of censoring weighting [23]; and landmarking [29,30], which scores only cohorts at least H days old with administrative censoring of follow-up at H, at H = 180, 365 and 730 days. The resulting trade-off between bias and delay is the known landmark-time selection trade-off [31]. The identifying assumption for weighting is tested directly rather than assumed: we fit a model for whether a label arrives within 365 days on the covariates a monitor can see, refit it with the true outcome added, and report the residual outcome dependence after covariate adjustment. This is testable here, although it is not testable in general [32], because the outcome is observed retrospectively even where the monitor could not see it.

Use of large language models

A large language model (Claude, Anthropic) was used in preparing this work: to draft and revise the analysis code supplied as Supplementary Information, and to draft and edit the manuscript text. All decisions concerning the research question, study design, data selection, outcome definitions, analytical

methods and experimental controls were the author's. The author independently verified every source, data item and result, and takes full responsibility for the accuracy and integrity of the manuscript. No large language model is listed as an author, and none could be: an LLM cannot satisfy authorship criteria and cannot be accountable for the work. Verification is mechanical as well as editorial. Every number printed in this manuscript is read from a result file produced by the analysis scripts, and a verification script supplied with them re-checks each printed figure against the file that produced it and fails on any disagreement.

Declarations

Ethics approval and consent

Not applicable. No human participants were enrolled and no identifiable private information was collected; all data are public US federal regulatory records. Most inspected entities are firms, but FDA's bioresearch-monitoring program inspects clinical investigators, so the source records name individuals as well as companies. No name is used as a feature, no name appears in this manuscript, and no name is included in the Supplementary Information. No consent was required.

Data Availability

All source data are public US federal records, and FDA states that the contents of its website are not copyrighted and are in the public domain. The two FDA Data Dashboard endpoints used are served through a credentialed API that requires a key issued under FDA's Unified Logon and that returns no export-date field, so a third party without a key cannot reproduce the retrieval; the fetch timestamp and SHA-256 digest of every retrieved file are recorded in the manifest supplied as Supplementary Information. The analysis code, every result file from which the numbers in this manuscript are read, and the figure code are supplied as Supplementary Information with this submission, together with the verification script that binds each printed number to the file that produced it. The random seed is 20260907 throughout. At the time of submission no public code repository and no persistent archive identifier exist for this work, and none is claimed; a public repository and a persistent identifier will be supplied at revision.

Author contributions

P.K.B. is the sole author and is responsible for all aspects of the work.

Additional Information

Competing interests. The author declares no competing interests.

Funding. This work received no funding.

Disclaimer. This work was carried out independently, on personal time and equipment, and is not connected to the author's employment. The views expressed are the author's own and do not represent the views, positions or policies of any current, former or future employer or client. No proprietary, confidential or internal data of any organization was used. All data is public: the FDA Data Dashboard inspection classification and cited observation endpoints (inspections_classifications, 341,046 records, and inspections_citations, 280,130 records, both fetched 7 September 2026), US federal public-domain records about corporate entities served through a credentialed API.

References

1. U.S. Food and Drug Administration. Inspection classifications. Office of Inspections and Investigations; NAI, VAI and OAI definitions <https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-basics/inspection-classifications> (2026).
2. Bhatt, J. A., Morris, K. R. & Haware, R. V. Development of predictive statistical model for gaining valuable insights in pharmaceutical product recalls. *AAPS PharmSciTech* **25**, 255; 10.1208/s12249-024-02970-z (2024).
3. Mulla, S. & Patel, M. Harnessing the power of artificial intelligence (AI) for insights into FDA recall patterns: a retrospective analysis. *AI Soc.* **41**, 1063-1074; 10.1007/s00146-025-02598-y (2026).
4. McGuire, C., Sellers, J. W. & Muldowney, L. A 7-year analysis of the U.S. FDA good clinical practice inspection outcomes for marketing applications. *Ther. Innov. Regul. Sci.* **59**, 1421-1431; 10.1007/s43441-025-00835-6 (2025).
5. Dmour, I. Content analysis of US FDA warning letters issued to compounding pharmacies regarding violations of current good manufacturing practices between 2017 and 2022. *J. Pharm. Innov.* **18**, 965-979; 10.1007/s12247-022-09692-4 (2023).
6. Kwiecinski, G. An analysis of FDA warning letter citations from 2019-2023. *J. Pharm. Innov.* **19**, 78; 10.1007/s12247-024-09879-x (2024).
7. Leung, K.-M., Elashoff, R. M. & Afifi, A. A. Censoring issues in survival analysis. *Annu. Rev. Public Health* **18**, 83-104; 10.1146/annurev.publhealth.18.1.83 (1997).
8. Kalbfleisch, J. D. & Prentice, R. L. *The Statistical Analysis of Failure Time Data* 2nd edn <https://doi.org/10.1002/9781118032985> (Wiley, 2002).
9. Begg, C. B. & Greenes, R. A. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* **39**, 207-215; 10.2307/2530820 (1983).
10. Zhou, X.-H. Correcting for verification bias in studies of a diagnostic test's accuracy. *Stat. Methods Med. Res.* **7**, 337-353; 10.1177/096228029800700403 (1998).
11. Tsiatis, A. A. A nonidentifiability aspect of the problem of competing risks. *Proc. Natl Acad. Sci. USA* **72**, 20-22; 10.1073/pnas.72.1.20 (1975).
12. Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Stat. Assoc.* **94**, 1096-1120;

- 10.1080/01621459.1999.10473862 (1999).
13. Arifin, W. N. & Yusof, U. K. Correcting for partial verification bias in diagnostic accuracy studies: a tutorial using R. *Stat. Med.* **41**, 1709-1727; 10.1002/sim.9311 (2022).
 14. Marrs, G. R., Hickey, R. J. & Black, M. M. The impact of latency on online classification learning with concept drift in *Knowledge Science, Engineering and Management (KSEM 2010)* (Lecture Notes in Computer Science vol. 6291) 459-469 https://doi.org/10.1007/978-3-642-15280-1_42 (Springer, 2010).
 15. Pham, T., Kottke, D., Krempel, G. & Sick, B. Stream-based active learning for sliding windows under the influence of verification latency. *Mach. Learn.* **111**, 2011-2036; 10.1007/s10994-021-06099-z (2022).
 16. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.* **46**, 1-37; 10.1145/2523813 (2014).
 17. Grzenda, M., Gomes, H. M. & Bifet, A. Delayed labelling evaluation for data streams. *Data Min. Knowl. Discov.* **34**, 1237-1266; 10.1007/s10618-019-00654-y (2020).
 18. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. & Bontempi, G. Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 3784-3797; 10.1109/TNNLS.2017.2736643 (2018).
 19. Souza, V. M. A., dos Reis, D. M., Maletzke, A. G. & Batista, G. E. A. P. A. Challenges in benchmarking stream learning algorithms with real-world data. *Data Min. Knowl. Discov.* **34**, 1805-1858; 10.1007/s10618-020-00698-5 (2020).
 20. Chapelle, O. Modeling delayed feedback in display advertising in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1097-1105 <https://doi.org/10.1145/2623330.2623634> (ACM, 2014).
 21. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D. & Matheny, M. E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Inform. Assoc.* **24**, 1052-1061; 10.1093/jamia/ocx030 (2017).
 22. Kore, A. et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat. Commun.* **15**, 1887; 10.1038/s41467-024-46142-w (2024).
 23. Robins, J. M. & Rotnitzky, A. Recovery of information and adjustment for dependent censoring using surrogate markers in *AIDS Epidemiology: Methodological Issues* 297-331 https://doi.org/10.1007/978-1-4757-1229-2_14 (Birkhäuser, 1992).
 24. Willems, S. J. W., Schat, A., van Noorden, M. S. & Fiocco, M. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Stat. Methods Med. Res.* **27**, 323-335; 10.1177/0962280216628900 (2018).
 25. European Medicines Agency. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023; adopted; first published 30 September 2024 <https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle> (2024).
 26. U.S. Food and Drug Administration. FDA Data Dashboard: inspections classifications and inspections citations. Credentialed API, api-datadashboard.fda.gov/v1; retrieved 7 September 2026

<https://datadashboard.fda.gov/> (2026).

27. Balagam, P. K. Predicting the inspector, not the plant: regulatory targeting and explanation instability in pharmaceutical inspection-outcome models. Preprint at <https://doi.org/10.21203/rs.3.rs-11022124/v1> (2026).
28. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283-286; 10.1056/NEJMc2104626 (2021).
29. Anderson, J. R., Cain, K. C. & Gelber, R. D. Analysis of survival by tumor response. *J. Clin. Oncol.* **1**, 710-719; 10.1200/JCO.1983.1.11.710 (1983).
30. van Houwelingen, H. C. Dynamic prediction by landmarking in event history analysis. *Scand. J. Stat.* **34**, 70-85; 10.1111/j.1467-9469.2006.00529.x (2007).
31. Dafni, U. Landmark analysis at the 25-year landmark point. *Circ. Cardiovasc. Qual. Outcomes* **4**, 363-371; 10.1161/CIRCOUTCOMES.110.957951 (2011).
32. Ying, A. Proximal survival analysis to handle dependent right censoring. *J. R. Stat. Soc. B* **86**, 1414-1434; 10.1093/jrsssb/qkae037 (2024).

Figures

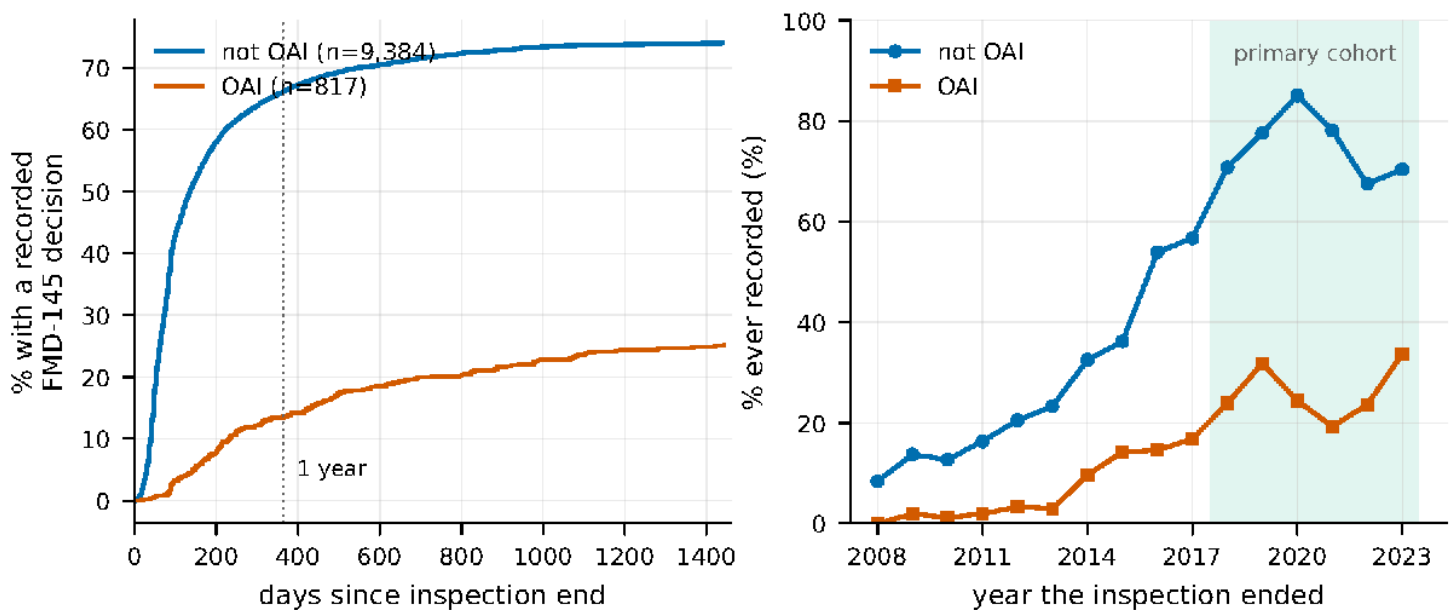


Figure 1

Left: cumulative share of inspections with a recorded FMD-145 decision, Kaplan-Meier with administrative censoring at the retrieval date, primary cohort; passing inspections reach 74% and do so mostly within six months, while failing inspections plateau near 25%. Right: share ever recorded by the year the inspection ended, with three-year follow-up; the upward trend is the export becoming more complete, and the gap between the two series is the effect this paper is about, which persists throughout.

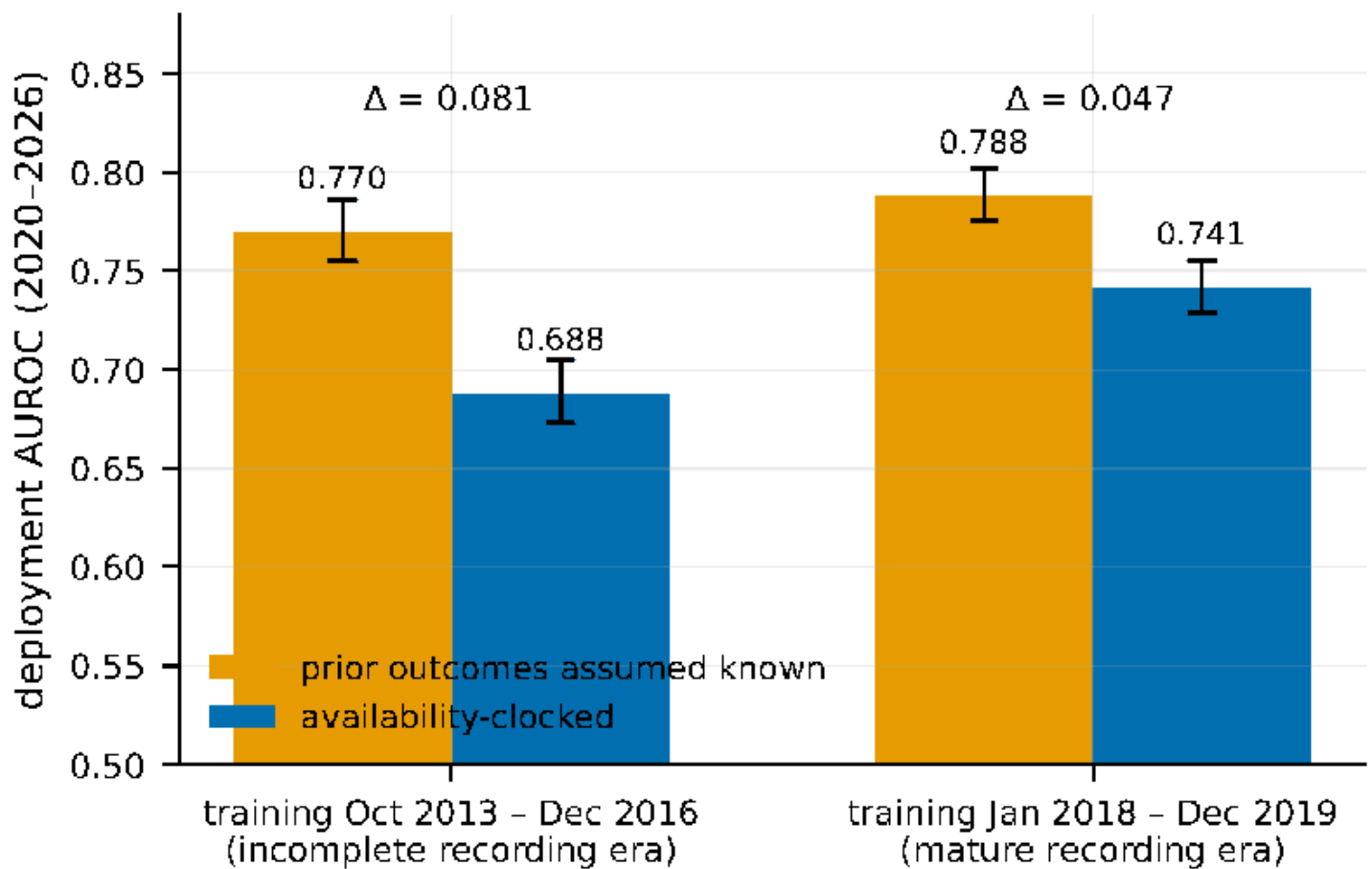


Figure 2

Deployment AUROC over 1 January 2020 to 31 May 2026, with and without an availability clock on the history features, for two training windows: October 2013 to December 2016, when the export is materially incomplete, and January 2018 to December 2019, when recording is mature. Error bars are 95% facility-level bootstrap intervals over 300 resamples, and the intervals do not overlap in either era.

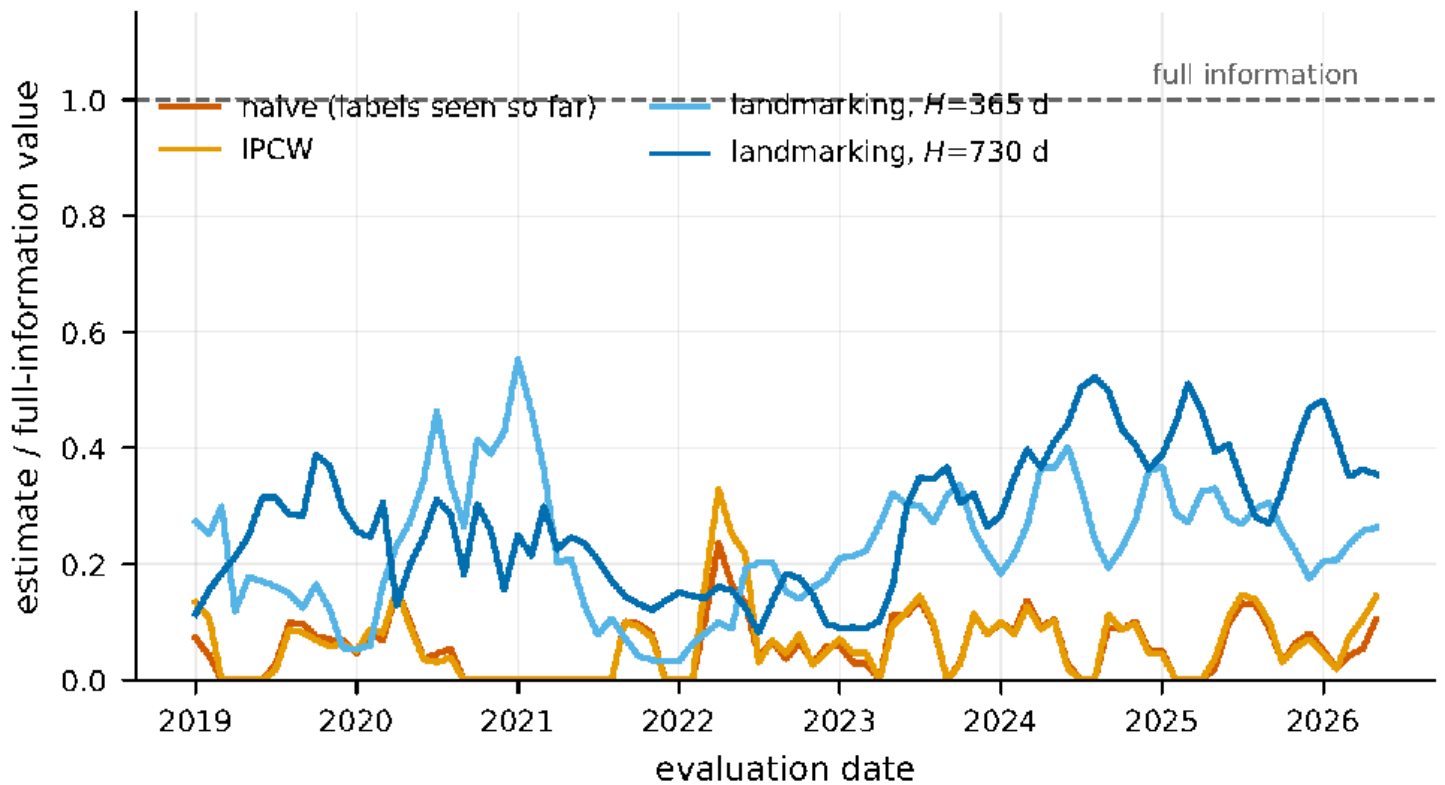


Figure 3

The naive, weighted and two landmark estimators, each divided by the full-information value on the same window; landmarking at $H = 180$ d is reported in Table 2 and omitted here for legibility. The naive and weighted series are almost indistinguishable, which is what a violated identifying assumption looks like in practice; landmarking separates from them and still runs at a quarter to a half of the truth.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.zip](#)