

Additional file 1

Supplementary methods, tables, and figures for

When does validation design change the answer? A simulation study and two empirical applications in clustered longitudinal health data

1 Data-integrity audit

Table S1: Source rows and effective report-level sample sizes

Characteristic	1996–2019	2020–2024
Country-level rows	3,338	301
Unique DON identifiers	2,789	298
Rows belonging to duplicated DON identifiers	777	6
Exact duplicated rows	0	0
Category A rows	593	74
Category A reports	459	73
Missing case-count rows	1,788 (53.6%)	172 (57.1%)
Date range	22 Jan 1996–26 Dec 2019	2 Jan 2020–27 Dec 2024

Country-level rows must be distinguished from unique DON reports: treating the 3,338 rows and 301 rows as independent events overstates the effective sample size. No exact duplicate rows were found. Repeated identifiers generally represent multi-country rows derived from one DON report. Because the later-period file also contained three repeated identifiers, its report-level denominator is 298 rather than 301.

Case counts were missing in more than half of both files. Missing counts are converted to zero in the primary analysis, following the most common default. The audit treats this as a modelling assumption and reports missingness-aware sensitivity analyses below. Future work should distinguish confirmed zero from unavailable information at data collection.

2 Feature-construction rules

2.1 Geographic Clustering Index

A symmetric ± 30 -day window would allow future records to alter the score of an earlier report. The past-only definition used here is

$$\text{GCI}_i = \frac{1}{|\{c : \text{disease}(c) = d_i, t_i - 30 \leq t_c \leq t_i\}|}.$$

Same-day rows are included because all countries named in the same released DON are available at publication time. No record dated after t_i is used.

2.2 Seasonal Mismatch Anomaly

Computing disease-month means and SDs before cross-validation would let evaluation-fold information enter the reference. The pipeline used here fits the SMA reference on each training subset.

Disease–month cells with fewer than five training rows use the corresponding training month reference; the global training reference is the final fallback. Strict leave-one-pathogen-out therefore does not use the held pathogen’s own seasonal distribution.

2.3 Evaluation grouping

All country rows sharing a DON identifier remain in one cross-validation fold. Predictions are aggregated to report level using the maximum row probability, representing an “alert if any country row triggers” policy. Only 228 of 2,789 reports (8.2%) contain more than one row, so the aggregation rule can affect only a small subset. Re-running the released pipeline with mean instead of maximum aggregation changed discrimination negligibly: the grouped cross-validation and later-period validation AUCs for M-SS each shifted by at most about 0.5 percentage points under mean rather than maximum aggregation. The choice of aggregation therefore does not alter any conclusion.

3 Seed-specific grouped cross-validation

Table S2: Five-fold StratifiedGroupKFold results at report level

Model	Seed	AUC	Sensitivity	Specificity	PPV
M-SS	42	0.9352	0.8998	0.8240	0.5018
M-SS	123	0.9352	0.9020	0.8318	0.5136
M-SS	456	0.9358	0.8932	0.8343	0.5151
M-SS	789	0.9380	0.8998	0.8283	0.5080
M-SS	2024	0.9307	0.8998	0.8236	0.5012
M-Year	42	0.9616	0.9041	0.8820	0.6014
M-Year	123	0.9620	0.8889	0.8914	0.6172
M-Year	456	0.9643	0.9150	0.8815	0.6034
M-Year	789	0.9610	0.8976	0.8858	0.6077
M-Year	2024	0.9622	0.9041	0.8833	0.6041

Each seed produced one out-of-fold prediction for all 2,789 reports, including 459 Category A reports. Threshold metrics use $P \geq 0.50$.

4 Seed-specific temporal holdout

The temporal holdout trained on 2,697 rows dated through 2014 and tested on 641 rows from 2015–2019. Test rows aggregated to 592 reports, including 120 Category A reports.

5 Strict leave-one-pathogen-out validation

The held pathogen was absent from model fitting and the SMA reference. Values are threshold-dependent. The very small Tularemia and Anthrax report counts make their estimates unstable. A non-strict holdout mixing epidemic- and pathogen-level exclusions gives 27.3% instead because it combined a West Africa Ebola epidemic holdout with whole-pathogen holdouts for other agents.

6 Later-period temporal validation and uncertainty

One thousand bootstrap replicates sampled 137 disease–time clusters. M-SS AUC was 0.831 (95% interval 0.741–0.884), sensitivity 0.795 (0.714–0.884), specificity 0.640 (0.550–0.717), and PPV 0.417 (bootstrap interval approximately 0.198–0.602). Paired M-SS minus M-Year differences were: AUC

Table S3: Temporal holdout results, 2015–2019

Model	Seed	AUC	Sensitivity	Specificity	PPV
M-SS	42	0.6160	0.4917	0.8581	0.4683
M-SS	123	0.6215	0.4917	0.8411	0.4403
M-SS	456	0.6191	0.5000	0.8475	0.4545
M-SS	789	0.6231	0.4917	0.8644	0.4797
M-SS	2024	0.6032	0.4833	0.8559	0.4603
M-Year	42	0.8109	0.6833	0.7775	0.4385
M-Year	123	0.8099	0.6083	0.7903	0.4244
M-Year	456	0.7947	0.4583	0.7903	0.3571
M-Year	789	0.8017	0.5500	0.7754	0.3837
M-Year	2024	0.8019	0.6667	0.7648	0.4188

Table S4: Strict LOPO detection at report level, mean $\pm$ SD across five seeds

Held pathogen	Rows	Reports	M-SS detection	M-Year detection
Ebola virus	436	308	0.383 $\pm$ 0.019	0.309 $\pm$ 0.021
Marburg fever	54	52	0.677 $\pm$ 0.066	0.573 $\pm$ 0.079
Lassa fever	39	37	0.411 $\pm$ 0.023	0.362 $\pm$ 0.015
Plague	37	36	0.722 $\pm$ 0.028	0.606 $\pm$ 0.072
Anthrax	18	18	0.289 $\pm$ 0.025	0.733 $\pm$ 0.256
Tularemia	5	5	0.520 $\pm$ 0.110	0.280 $\pm$ 0.179
Macro mean			0.500	0.477

0.034 (−0.014 to 0.101), sensitivity 0.055 (−0.030 to 0.182), specificity −0.031 (−0.096 to 0.025), and PPV −0.005 (−0.056 to 0.046).

7 Calibration, decision-curve, and robustness stress tests

Both specifications overpredicted the proxy outcome prevalence in the later-period temporal validation set. Calibration-in-the-large was therefore inadequate for direct resource allocation, even though discrimination was moderate and Brier scores were similar.

Adding a missingness indicator did not reduce temporal-validation performance, whereas replacing missing counts with the training median degraded performance. This pattern supports the manuscript’s interpretation that missingness is informative and must be treated as part of the validation problem.

Temporal-block validation was intentionally interpreted as a heterogeneity stress test rather than as a replacement for prospective validation. The low 2010–2014 AUCs indicate a major calendar-period shift and show why a single grouped cross-validation estimate is not sufficient. The 2015–2019 block uses the same development and test split as the primary temporal holdout (main-text Table 2) and is therefore numerically identical to it by construction. All blocks were computed with the single, common past-only GCI construction used in the main analysis, so every validation contrast reported here and in the main text is mutually consistent.

Because the endpoint is a methodological proxy, decision-curve analysis was used only to show threshold dependence. It should not be read as evidence that either model is ready for operational deployment.

Table S5: Report-level confusion matrices, 2020–2024

Model	TP	FN	TN	FP	Sens.	Spec.	PPV	Brier
M-SS	58	15	144	81	0.795	0.640	0.417	0.196
M-Year	54	19	151	74	0.740	0.671	0.422	0.197

Table S6: Calibration summaries in the 2020–2024 temporal validation set

Model	Brier	Mean predicted	Observed	O:E	Calibration slope
M-SS	0.196	0.464	0.245	0.528	1.085
M-Year	0.197	0.425	0.245	0.577	0.698

8 Cascade scenarios

The cascade was implemented prospectively in time: an alert could propagate only to later reports. S1/S2 recovered 11 additional positive reports but produced 12 additional false positives. Nomenclature normalisation did not alter the aggregate confusion matrix in this report-level analysis.

9 Reproducibility map

- `code/validate_and_reanalyse.py`: data audit, causal GCI, fold-fitted SMA, grouped CV, temporal holdout, strict LOPO, later-period temporal validation, cascade.
- `code/bootstrap_external.py`: disease–time cluster bootstrap for the later-period temporal validation set.
- `code/make_verified_figures.py`: all main-text figures.
- `stress_test_analyses.py`: calibration summaries, decision-curve illustration, case-count missingness sensitivity, and temporal-block internal–external validation.
- `verified_results/verified_results.json`: machine-readable metrics.
- `verified_results/temporal_predictions_verified.csv`: report-level temporal-validation predictions.
- `verified_results/temporal_cluster_bootstrap.json`: uncertainty intervals.
- `stress_test_analyses.json`: machine-readable stress-test outputs for the additional analyses.
- `sim/`: controlled-simulation code (factorial ICC×drift grid, outcome-concordance and model-family robustness experiments) and cached results.
- `ehr/`: hospital-readmission analysis code and summary for the second dataset.
- `leakage_diagnostic.py`: released pre-modelling diagnostic (feature ICC, label ICC, temporal shift → expected leakage and temporal-optimism regimes).

10 Model-family robustness of the simulation

The two governing effects were re-derived with two additional learners on the simulation data (three seeds; leakage slice at drift $\delta = 0$, drift slice at fixed $\sigma_a = 1.5$). Temporal optimism under drift is model-agnostic, whereas the row-level leakage gap is realised only by sufficiently flexible learners.

Random forests (main text) behaved like gradient boosting. The near-zero leakage gap for logistic regression is expected: a linear model cannot memorise the cluster-identity nuisance features that carry the leak, so row grouping changes little. Concept-drift optimism, by contrast, appears for every model family because it reflects a genuine train–test distribution shift rather than memorisation.

Table S7: Case-count missingness sensitivity for M-SS in 2020–2024

Specification	AUC	Brier	Sens.	Spec.	PPV	TP	FP	FN
Zero imputation	0.831	0.196	0.795	0.640	0.417	58	81	15
Zero + missing indicator	0.836	0.193	0.822	0.653	0.435	60	78	13
Median + missing indicator	0.748	0.219	0.740	0.582	0.365	54	94	19

Table S8: Temporal-block internal–external validation, AUC mean $\pm$ SD across five seeds

Held-out period	Reports	Positive reports	M-SS AUC	M-Year AUC
1996–2004	1,009	192	0.808 $\pm$ 0.010	0.835 $\pm$ 0.008
2005–2009	624	57	0.946 $\pm$ 0.002	0.948 $\pm$ 0.003
2010–2014	564	90	0.464 $\pm$ 0.017	0.479 $\pm$ 0.017
2015–2019	592	120	0.617 $\pm$ 0.008	0.804 $\pm$ 0.007

A lightweight diagnostic implementing these checks (`leakage_diagnostic.py`) is released with the code: given a table with a cluster identifier, optional time column, features, and label, it returns the feature ICC, the label ICC, and a temporal label shift, and reports the expected leakage and temporal-optimism regimes before a model is trained.

11 Supplementary figures

Table S9: Selected decision-curve net-benefit values for M-SS

Risk threshold	0.10	0.20	0.50
Net benefit	0.167	0.109	−0.077

Table S10: Sensitivity–specificity tradeoff of cascade processing

Scenario	TP	FN	TN	FP	Sensitivity	Specificity	PPV
S0, no cascade	58	15	144	81	0.795	0.640	0.417
S1, forward 30-day cascade	69	4	132	93	0.945	0.587	0.426
S2, cascade + taxonomy	69	4	132	93	0.945	0.587	0.426

Table S11: Leakage gap (naive – grouped CV) and optimism gap (grouped – temporal), in AUROC percentage points, by model family.

Gap	Model	low	mid	high
Leakage gap (by within-cluster correlation)	Logistic regression	−0.2	−0.2	−0.1
	Gradient boosting	0.7	3.8	4.3
Optimism gap (by concept drift)	Logistic regression	−1.5	8.9	19.8
	Gradient boosting	−1.8	6.5	20.1

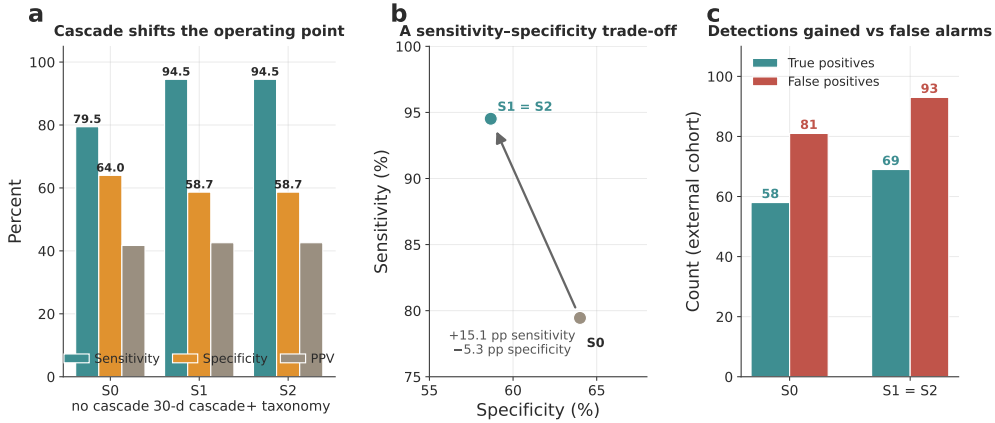


Figure S1: **Cascade propagation is a sensitivity–specificity trade-off (surveillance case study).** **a**, Sensitivity, specificity, and PPV for individual report scoring (S0), a forward-only 30-day cascade (S1), and cascade plus nomenclature normalisation (S2); S2 equalled S1. **b**, Movement in specificity–sensitivity space from S0 to S1/S2 (+15.1 sensitivity, −5.3 specificity points). **c**, Absolute counts: the cascade added 11 true positives (58→69) at the cost of 12 false positives (81→93).

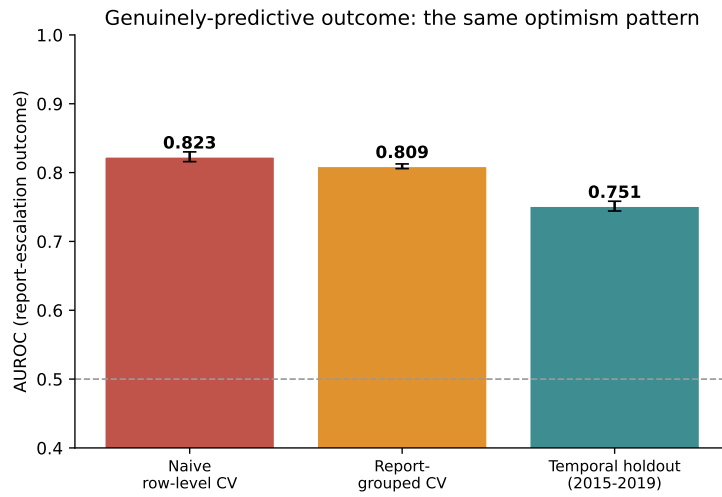


Figure S2: **The same optimism pattern for a genuinely predictive outcome.** AUROC for report escalation (same-disease follow-up within 90 days) under naive row-level cross-validation, report-grouped cross-validation, and 2015–2019 temporal holdout, using the M-SS feature set (mean±SD across five seeds).

12 TRIPOD+AI reporting checklist

The table below records, for every item of the TRIPOD+AI checklist, where in the main manuscript that item is addressed. Item numbers, topic labels, and the development/evaluation applicability markers follow the TRIPOD+AI statement (Collins GS, Moons KGM, Dhiman P, et al. BMJ. 2024;385:e078378; <https://www.tripod-statement.org>). The full wording of each checklist item is not reproduced here; readers should consult the official checklist. **D** denotes items relevant only to model development, **E** items relevant only to model evaluation, and **D;E** items applicable to both.

Several items are marked *not applicable*. This study compares validation designs; the two model specifications are methodological test vehicles and are not proposed for clinical or operational use. Items concerned with fairness across sociodemographic groups, patient and public involvement, model updating, and end-user interaction therefore have no substantive content in this study, and we state this explicitly rather than leaving them blank. Item 12e (heterogeneity across clusters) is not an ancillary consideration here but the central question of the study; TRIPOD-Cluster considerations apply throughout.

Section/topic	Item	D/E	Reported in	Note
Title	1	D;E	Title page (p. 1)	
Abstract	2	D;E	Abstract (p. 1–2)	
Background	3a	D;E	Background, para. 1–2 (p. 3)	
	3b	D;E	Background, para. 2 and 6 (p. 3–5)	Care-pathway context; both applications are methodological test beds, stated explicitly.
	3c	D;E	Not applicable	No sociodemographic predictors or protected attributes are used.
Objectives	4	D;E	Background, final three para. (p. 4–5)	The study evaluates validation designs; it does not propose a model for use.
Data	5a	D;E	Methods — First empirical application (p. 9); Second empirical application (p. 12)	
	5b	D;E	Methods — First empirical application (p. 9); Validation regimes (p. 12)	Development and evaluation periods defined per validation regime.
Participants	6a	D;E	Methods — Data-integrity rules (p. 10); Results — Cohort integrity (p. 16)	
	6b	D;E	Methods — First and Second empirical application (p. 9, 12)	WHO global reporting system; 130 US hospitals.
	6c	D;E	Methods — Data-integrity rules and evaluation unit (p. 10)	
Data preparation	7	D;E	Methods — Data-integrity rules (p. 10); Leakage-aware feature construction (p. 11)	Fold-internal fitting of data-derived transformations is a primary object of study.
Outcome	8a	D;E	Methods — First empirical application (p. 9); Table 1 (p. 10)	
	8b	D;E	Methods — First empirical application (p. 9); Results — Transportability (p. 22)	Primary outcome is a deterministic proxy; the secondary outcome uses a 90-day horizon.
	8c	D;E	Not applicable	Outcomes are deterministic or administratively recorded.
Predictors	9a	D	Methods — Leakage-aware feature construction (p. 11)	
	9b	D;E	Methods — Leakage-aware feature construction (p. 11); Model specification (p. 11)	Predictor sets fixed a priori to isolate the effect of validation design.
	9c	D;E	Methods — Leakage-aware feature construction (p. 11)	Prediction time and admissible information specified in Table 1 (p. 10).

Section/topic	Item	D/E	Reported in	Note
Sample size	10	D;E	Results — Cohort integrity and effective sample size (p. 16); Methods — Statistical analysis (p. 13)	Effective report-level sample size distinguished from row count.
Missing data	11	D;E	Methods — Data-integrity rules (p. 10); Results — stress tests (p. 21); Table S7	Three missingness strategies compared.
Analytical methods	12a	D	Methods — Model specification (p. 11); Validation regimes (p. 12); Statistical analysis (p. 13)	
	12b	D	Methods — Study design and overview (p. 5); Validation regimes (p. 12)	Four partitioning schemes compared (naive row-level, grouped, temporal hold-out, temporal blocks); this comparison is the study’s object.
	12c	D	Methods — Leakage-aware feature construction (p. 11)	
	12d	D;E	Methods — Model specification (p. 11); Validation regimes (p. 12)	Hyperparameters fixed, not tuned; stated as a limitation (p. 29).
	12e	D;E	Methods — Pre-modelling diagnostics (p. 8), Validation regimes (p. 12), Statistical analysis (p. 13); Results — Diagnostic coordinates (p. 24) and Table 3 (p. 25); temporal-block analysis (p. 21)	Heterogeneity across clusters and calendar blocks is the central question of this study. Reported also against TRIPOD-Cluster, which is cited in the main text.
	12f	E	Methods — Statistical analysis (p. 13)	Discrimination, calibration, and decision-curve measures pre-specified.
Class imbalance	13	D;E	Methods — Model specification (p. 11)	Class weighting set to the training negative-to-positive ratio.
Fairness	14	D;E	Not applicable (see note)	No sociodemographic or protected attributes are used and no subgroup fairness claim is made. Performance heterogeneity is instead reported across pathogens (p. 18) and calendar blocks (p. 21).
Model output	15	D	Methods — Model specification (p. 11); Table 1, “Intended interpretation” (p. 10)	Output is a ranking/triage score, explicitly not a diagnosis.
Training versus evaluation	16	D;E	Methods — Study design and overview (p. 5); Validation regimes (p. 12)	Separation of training from evaluation data is the study’s primary subject.
Ethical approval	17	D;E	Declarations — Ethics approval and consent to participate (p. 34)	
Funding	18a	D;E	Declarations — Funding (p. 35)	
Conflicts of interest	18b	D;E	Declarations — Competing interests (p. 35)	
Protocol	18c	D;E	No protocol was prepared	Secondary analysis of public data; stated for transparency.
Registration	18d	D;E	Not registered	Methodological study using public data.
Data sharing	18e	D;E	Declarations — Availability of data and materials (p. 34)	Raw snapshots released.
Code sharing	18f	D;E	Declarations — Availability of data and materials (p. 34)	Analysis code, figure scripts, seeds, and the released diagnostic utility.
Patient and public involvement	19	D;E	Not applicable	No patients or members of the public were involved; aggregate public reports and a de-identified public hospital dataset were used.
Participants	20a	D;E	Results — Cohort integrity and effective sample size (p. 16)	
	20b	D;E	Results — Cohort integrity (p. 16); Table S1	Row-to-report reduction and outcome counts reported for both periods.
	20c	E	Results — Cohort integrity (p. 16); Table 3 (p. 25); Table S1	Key dates, outcome frequency, missing-data volume, and cluster structure reported per dataset.
Model development	21	D;E	Methods — Model specification (p. 11); Results (p. 17)	

Section/topic	Item	D/E	Reported in	Note
Model specification	22	D	Methods — Model specification (p. 11); code repository	Reproducible from released code and fixed seeds.
Model performance	23a	D;E	Results — Table 2 (p. 18); Figs. 2–5	
	23b	D;E	Results — later-period validation (p. 19); LOPO (p. 18)	Cluster-bootstrap 95% intervals reported for all comparisons; per-pathogen performance reported separately.
Model updating	24	E	Not applicable	No model updating was performed.
Interpretation	25	D;E	Discussion — Principal findings (p. 25); Comparison with existing literature (p. 28)	
Limitations	26	D;E	Discussion — Strengths and limitations (p. 29)	A simulation-repetition caveat and five study limitations are stated.
Usability	27a	D	Methods — Data-integrity rules (p. 10); Results — missingness sensitivity (p. 21)	
	27b	D	Not applicable	The models are methodological test vehicles and are not proposed for deployment (Table 1, p. 10; Conclusions, p. 32).
	27c	D;E	Discussion — Recommendations (p. 31); Conclusions (p. 32)	