

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

**HanYeol Nam**

Ministry of National Defense

**Jin Yoo**

Ministry of National Defense

**Doo-Hee Lee**

Ministry of National Defense

**Jeongyun Kim**

Seoul National University

**Yoon Jeong Jang**

Ministry of National Defense

**Ku Kang**

**bisu9082@gmail.com**

Ministry of National Defense

---

## Research Article

**Keywords:** Clinical prediction models, Model validation, Optimism bias, Clustered data, Intraclass correlation, Cross-validation, Temporal validation, Reproducibility of results

**Posted Date:** October 6th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10788572/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

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

HanYeol Nam<sup>1</sup>, Jin Yoo<sup>1</sup>, Doo-Hee Lee<sup>1</sup>, Jeongyun Kim<sup>2</sup>, Yoon Jeong Jang<sup>1</sup>,  
and Ku Kang<sup>1,\*</sup>

<sup>1</sup>CBRN Defense Research Institute, Korea Ministry of National Defense,  
Seoul, Republic of Korea

<sup>2</sup>Department of Chemical and Biological Engineering, Seoul National  
University, Seoul, Republic of Korea

\*Correspondence: Ku Kang, bisu9082@gmail.com

**Author email addresses and ORCID iDs.** HanYeol Nam, namhy915@gmail.com, ORCID

0000-0001-7832-4687; Jin Yoo, tysonyj@snu.ac.kr, ORCID 0000-0002-7738-8424; Doo-Hee Lee,  
dooheeleechem@gmail.com, ORCID 0000-0002-5116-2064; Jeongyun Kim, kimjy3022@snu.ac.kr, ORCID  
0009-0004-1377-8572; Yoon Jeong Jang, yjjangchem@gmail.com, ORCID 0000-0002-5625-2528; Ku Kang,  
bisu9082@gmail.com, ORCID 0000-0001-8739-218X.

## Abstract

**Background.** Prediction models developed on clustered, longitudinal health data routinely report high internal discrimination that does not survive grouped or temporally separated evaluation. Guidance recommends subject- or record-aware resampling unconditionally, but does not indicate which datasets require it, or how much the reported performance changes if it is omitted. We asked which measurable data properties determine whether, and by how much, the choice of validation design changes a study's conclusions.

**Methods.** A controlled simulation generated clustered longitudinal datasets with independently tunable within-cluster feature correlation, outcome concordance, and concept drift, and compared naive row-level, grouped, and temporally separated validation. The same ladder, with pre-modelling diagnostics (feature and label intraclass correlation coefficients (ICCs), cluster size, and a drift contrast), was applied to a 24-year World Health Organization Disease Outbreak News archive (3,338 rows in 2,789 reports, 1996–2019; 298 reports, 2020–2024) and to 101,766 hospital encounters from 71,518 patients (30-day readmission).

**Results.** In simulation the row-level leakage gap reached 20.8 AUROC percentage points and the temporal gap 19.2 points; ICC governed the former and drift the latter, but they interacted: the leakage gap more than doubled with drift at fixed feature ICC (9.0 to 20.8 points). Neither empirical dataset reproduced a large leakage gap. The surveillance archive sat at the theoretical maximum of the ICC condition (label ICC 1.00, feature ICC 0.83), yet grouping changed AUC by only 0.8 percentage points because 92% of reports contributed a single row; the readmission cohort, at the opposite ICC corner (label ICC  $-0.02$ ), gave 0.1 points (0.2 among repeat admissions). In the archive, where a temporal split was available, that contrast dominated: grouped internal AUCs of 0.935 and 0.962 fell to 0.617 and 0.804 at temporal holdout.

**Conclusions.** Within-cluster correlation is necessary but not sufficient for row-level leakage. Cluster size is the binding constraint in real data: both datasets were leak-resistant despite occupying opposite corners of the ICC space. Effort spent enforcing grouped resampling may be better directed at temporal validation, and the diagnostics distinguishing the two regimes can be computed before a model is fitted. Data, code, and the diagnostic utility are openly released.

**Keywords:** Clinical prediction models; Model validation; Optimism bias; Clustered data; Intraclass correlation; Cross-validation; Temporal validation; Reproducibility of results

# 1 Background

Clinical prediction models are judged by their validation. Reported discrimination, calibration, and threshold performance are the evidence on which editors, reviewers, and health systems decide whether a model merits further evaluation, so the design of the validation study is part of the scientific claim rather than a downstream technical detail. Reporting and appraisal frameworks have converged on this position: TRIPOD and TRIPOD+AI require explicit statements of intended use, data provenance, missingness handling, calibration, and transportability, and PROBAST places resampling design squarely inside the analysis domain of risk-of-bias assessment[1, 2, 3, 4, 5, 6]. Methodological work on prediction models has likewise emphasised interpreting discrimination alongside calibration, accounting for sample-size uncertainty, and preferring internal–external over single-split validation[7, 8, 9, 10, 11, 12, 13, 14, 15], and recent guidance sets out explicitly how a model should progress from development to external validation before any claim of usefulness is made[16].

A recurring failure mode motivates this guidance. When observations are clustered — repeated encounters from one patient, repeated rows from one report, repeated records from one site — random row-level cross-validation places members of the same cluster in both the development and the evaluation set, and a sufficiently flexible learner can score them by recognising the cluster rather than by modelling the intended biomedical signal. The general phenomenon and the corresponding recommendation to split by subject or record are well documented[17, 18, 19, 20, 21, 22], as is the parallel observation that internal estimates degrade when site, acquisition, coding, or calendar period shifts between development and deployment[23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. Small effective sample sizes make the resulting estimates additionally unstable[34, 35, 36]. CONSORT-AI, SPIRIT-AI, DECIDE-AI, MI-CLAIM, and CLAIM address different study designs but converge on the same requirement: evaluation should approximate intended use and report uncertainty and failure modes rather than a single favourable point estimate[37, 38, 39, 40, 41, 42].

What this literature does not provide is a quantitative account of *when* the recommended validation choices change a study’s conclusions, and by *how much*. Grouped resampling is rec-

ommended universally, yet applying it is not free: it reduces the effective sample size, complicates nested tuning, and is sometimes omitted precisely because analysts cannot tell in advance whether it would matter. The converse error is equally consequential — reporting a grouped estimate as though it were transportable when the mechanism generating the outcome has drifted. Wynants et al. addressed part of this question, showing in a simulation of multicentre data that ignoring clustering alters both calibration and discrimination and that the intraclass correlation coefficient (ICC) is a governing quantity[43]; Little, Varoquaux and colleagues, in a joint commentary with the original authors, argued that the appropriate resampling scheme depends on the inference intended rather than being universally fixed[44]. What is still missing is a decomposition that separates the feature-side from the outcome-side component of clustering, adds the temporal axis, and—critically—is checked against real datasets to see whether the simulated magnitudes actually arise there. This is an empirical question about the association between a methodological choice and the study outcome it produces, and it is the question we address here.

Our starting hypothesis was that the direction and approximate magnitude of validation optimism reduce to three properties of the data, each estimable before training: the degree to which predictors carry cluster identity (feature ICC), the degree to which the outcome is shared within clusters (label ICC), and the degree to which the predictor–outcome relationship changes over time (concept drift). We expected the first two to govern row-level leakage jointly, so that either alone would be insufficient, and the third to govern the gap between internal and temporally separated estimates. As reported below, the simulation supported this hypothesis but the two empirical datasets did not reproduce the leakage magnitudes it implies, and reconciling the two required a fourth quantity—cluster size—that the ICC framing does not capture.

We tested this in three stages. A controlled simulation varied the three properties independently and measured the resulting gaps between naive row-level, cluster-grouped, and temporally separated validation. We then applied an identical validation ladder, and the corresponding diagnostics, to two empirical datasets that turn out to occupy opposite corners of the ICC space. We note the order of work explicitly, because it bears on how the agreement between simulation and data should be read: the surveillance archive was analysed first, the simulation was

built afterwards to identify the mechanism behind what we had observed, and the readmission cohort was then added as a contrasting test of the outcome-concordance condition. The two datasets were therefore not selected to span a previously-specified parameter space, and we report their measured diagnostic coordinates rather than asserting where they sit. The first is a 24-year archive of World Health Organization Disease Outbreak News (DON) reports, a stringent test case because a single report may generate several country-level rows, the same outbreak generates many reports, disease prevalence changes across calendar epochs, and case counts are frequently missing[45]; here the outcome is defined at the report level and is therefore highly concordant within cluster. The second is a public hospital-readmission cohort in which the natural cluster is the patient but readmission is largely independent across a patient’s admissions[46], placing it at the predicted null condition. Because a deterministic proxy label is used in the surveillance application, we also repeated the ladder on a genuinely predictive secondary outcome from the same archive to confirm that the pattern is not an artefact of the proxy.

Our aim is not to propose a new estimator or resampling scheme; every validation regime we compare is already established. The contribution is to quantify how much each choice changes the reported result, to identify the data conditions under which it does so, and to convert that finding into a diagnostic that can be run before modelling. We report the study against TRIPOD+AI and release the data, code, and diagnostic utility in full.

## 2 Methods

### 2.1 Study design and overview

The study has three components, which the Results present in logical rather than chronological order (the chronology is given in the Background). A controlled simulation establishes, under a known data-generating mechanism, how much three independently tunable data properties change the estimate produced by each validation design. A first-order approximation formalises the relationship the simulation tests and defines the null conditions under which the designs

should agree. The identical validation ladder, and the corresponding pre-modelling diagnostics, are then applied to two empirical datasets with contrasting cluster structure: a longitudinal surveillance archive with a report-level outcome, and a hospital cohort with an encounter-level outcome. Each Results subsection reports the contrast between validation designs rather than the performance of any single model: the object of study throughout is the difference between the numbers that established designs produce on the same data.

## 2.2 Controlled simulation

To characterise validation optimism independently of any real archive, we generated synthetic clustered longitudinal datasets with three independently tunable properties. The simulation is reported following the aims, data-generating mechanisms, estimands, methods, and performance measures (ADEMP) structure[47]; the estimand is the difference in report-level AUROC between two validation designs applied to the same generated dataset, and the performance measure is its mean over repetitions. Each dataset contained 300 clusters (analogous to reports); each cluster  $c$  was assigned a publication time  $t_c \sim \text{Uniform}(0, 1)$  and  $1 + \text{Poisson}(1.5)$  member rows. Four latent signal variables  $Z_c \in \mathbb{R}^4$  generated a binary cluster-level outcome through a time-varying linear predictor  $\eta_c = 2.2 b(t_c)^\top Z_c$ , where the coefficient vector rotated with time,  $b(t_c) = \cos(\theta_c) b_0 + \sin(\theta_c) b_\perp$  with  $\theta_c = \delta (t_c - 0.5)$  and  $b_\perp \perp b_0$ ; the parameter  $\delta$  therefore controls concept drift ( $\delta = 0$  gives a stationary mechanism). Observed features were the four signal variables (each  $Z_c$  plus Gaussian noise) and eight nuisance features carrying a cluster-identity vector  $a_c \sim \mathcal{N}(0, \sigma_a^2)$ ;  $\sigma_a$  therefore controls within-cluster correlation, because rows of the same cluster share  $a_c$  while the outcome does not depend on it, so memorising  $a_c$  helps only when the same cluster appears in both training and test folds (leakage). A random forest (80 trees, maximum depth 12) was trained and evaluated under three regimes: naive row-level  $k$ -fold cross-validation ( $k = 4$ ), report-grouped cross-validation (`StratifiedGroupKFold`,  $k = 4$ ), and a temporal holdout (train  $t_c < 0.7$ , test  $t_c \geq 0.7$ ). Predictions were aggregated to the cluster level and scored by AUROC. We swept  $\sigma_a \in \{0, 0.75, 1.5, 2.5, 4.0\}$  at  $\delta = 0$  (leakage slice) and  $\delta \in \{0, 0.75, 1.5, 2.5, 4.0\}$  at  $\sigma_a = 1.5$  (drift slice), averaging over four generated datasets per cell, and additionally ran the full  $5 \times 5$  factorial over both parameters (three

per cell). For interpretability, the nuisance correlation was summarised by the empirically measured one-way random-effects intraclass correlation coefficient (ICC(1)) of the nuisance features. In a separate experiment we varied within-cluster *outcome* concordance: each row’s label equalled its cluster label with probability  $\rho_y$  and was otherwise drawn independently at the marginal rate, with  $\rho_y \in \{0, 0.25, 0.5, 0.75, 1.0\}$  (fixed  $\sigma_a = 1.5$ ,  $\delta = 0$ ); rows were scored at the row level to mirror encounter-level clinical outcomes, and the resulting label ICC(1) was reported alongside the leakage gap.

## 2.3 Terminology

We use *validation optimism* as an umbrella term for the difference between what two validation designs report on the same data, and distinguish two components throughout.  $\Delta_{\text{row}}$ , the difference between naive row-level and cluster-grouped cross-validation, is clustering-induced optimism.  $\Delta_{\text{time}}$ , the difference between grouped internal validation and a temporally separated holdout, is a temporal transportability gap rather than optimism in the classical sense. Both are distinct from the apparent-minus-corrected optimism of Efron and of Harrell, which quantifies overfitting within a single population and is not the object of this study[36, 15, 8].

## 2.4 Diagnostic approximation for validation optimism

The simulation was designed to test a simple approximation rather than to define a single favourable data-generating mechanism. Let observations  $i$  be nested in clusters  $c(i)$ , let  $S_i$  denote the model score, and write the score as the sum of a transportable signal component and a cluster-memory component,

$$S_i = g(X_i) + m\{c(i)\},$$

where  $m\{c(i)\}$  can be learned only when the same, or a near-identical, cluster is represented in both development and validation data. In row-level cross-validation, the expected extra apparent discrimination contributed by  $m$  is negligible unless two quantities are both non-zero: the predictors must carry cluster identity, measured by a feature ICC  $\rho_X$ , and the outcome must

184 be concordant within cluster, measured by a label ICC  $\rho_Y$ . To first order,

$$\Delta_{\text{row}} = \text{AUC}_{\text{row CV}} - \text{AUC}_{\text{grouped CV}} \approx K_f \rho_X \rho_Y,$$

185 where  $K_f \geq 0$  is a learner-dependent flexibility constant. Linear or strongly regularised models  
 186 have small  $K_f$  because they cannot easily encode cluster identity from nuisance variation;  
 187 flexible tree ensembles have larger  $K_f$ . This approximation gives the qualitative null conditions  
 188 used in the audit: if either  $\rho_X \approx 0$  or  $\rho_Y \approx 0$ , row-level leakage should be small even when  
 189 repeated observations are present.

190 Temporal optimism is different. Grouped internal validation can still be optimistic when  
 191 the conditional relationship between predictors and outcome changes over time. Let  $P_t(Y | X)$   
 192 denote the prediction mechanism at time  $t$  and let  $D(t_0, t_1)$  be an empirical drift contrast, such  
 193 as a coefficient-rotation distance, calibration-slope change, or blockwise AUC heterogeneity.  
 194 The corresponding first-order audit is

$$\Delta_{\text{time}} = \text{AUC}_{\text{grouped CV}} - \text{AUC}_{\text{temporal}} \approx K_t D(t_{\text{train}}, t_{\text{test}}),$$

195 with  $K_t$  depending on model family, feature set, and evaluation window. The approximation  
 196 is not used as a formal estimator of AUC loss; it is a pre-modelling diagnostic that predicts  
 197 which validation contrast must be reported. The simulations varied  $\rho_X$ ,  $\rho_Y$ , and drift separately  
 198 to verify the null conditions and monotonicity implied by these expressions, and the two real  
 199 datasets tested the two boundary cases: high report-level outcome concordance in surveillance  
 200 data and near-zero patient-level outcome concordance in readmission data.

## 201 **2.5 Pre-modelling diagnostics computed on the empirical datasets**

202 For each empirical dataset we computed the quantities the approximation requires, using only  
 203 information available before a model is fitted.

204 *Feature and label ICC.* One-way random-effects intraclass correlation coefficients,  $\text{ICC}(1)$ ,  
 205 were computed for the outcome and for each predictor with respect to the natural cluster (the

DON report; the patient). Unbalanced cluster sizes were handled with the standard  $k_0$  correction, and estimates are reported unclamped, so that a negative value—indicating that within-cluster variance exceeds between-cluster variance—is shown as such rather than truncated to zero. In the surveillance archive, two predictors (the seasonal magnitude anomaly and its interaction with the geographic clustering index) are fitted inside training folds by construction and therefore have no pre-modelling column; ICCs are reported for the four persistent predictors.

*Cluster size.* Because ICC describes how strongly cluster membership predicts a value but not how often membership is shared, we also report the mean number of rows per cluster and the proportion of clusters contributing more than one row. A dataset in which almost every cluster is a singleton offers a learner little opportunity to exploit shared membership regardless of its ICC.

*Drift contrast.* The temporal approximation requires an observable stand-in for the simulation’s coefficient-rotation parameter  $\delta$ , which is not identifiable outside a known data-generating mechanism. We used the cosine distance between standardised logistic-regression coefficient vectors fitted separately to the development and the evaluation period,  $D = 1 - \cos(\hat{\beta}_{\text{dev}}, \hat{\beta}_{\text{eval}})$ , so that  $D = 0$  denotes an unchanged predictor–outcome relationship and  $D = 1$  orthogonal ones. The same definition was applied to both temporal contrasts in the surveillance archive.

## 2.6 First empirical application: WHO Disease Outbreak News archive

Training data were the public retrospective WHO DON database for 1996–2019[45]. Later-period data were collected from the WHO DON API for 2020–2024. The analysis used public, non-individual, aggregate outbreak reports; no personal health information was processed. The 2020–2024 data are termed a *later-period validation set*, not a prospective cohort, because they were retrospectively assembled from the same WHO reporting source; this terminology is used consistently throughout to avoid implying independent prospective collection.

The binary methodological proxy was whether `DiseaseLevel1` matched Ebola virus, Marburg fever, Lassa fever, plague, anthrax, tularemia, botulism, or smallpox, following the United States Centers for Disease Control and Prevention bioterrorism agent Category A list

(<https://emergency.cdc.gov/agent/agentlist-category.asp>, accessed 12 May 2026). This label tests validation behaviour in a temporally clustered archive; it is not proposed as a clinical diagnosis or a substitute for deterministic disease-name mapping. Table 1 states the target estimand, the prediction time, and the role each validation design plays.

**Table 1.** Target estimand and validation role

| Component               | Specification                                                                                                                                                         |
|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Evaluation unit         | DON report, not country-level row; multi-country report probabilities aggregated by maximum row-level probability.                                                    |
| Prediction time         | Date of report publication, using only information available on or before that date.                                                                                  |
| Target population       | Public WHO DON reports in the same reporting ecosystem; 2020–2024 data are later-period temporal validation, not independent prospective validation.                  |
| Outcome role            | CDC Category A match used as a reproducible methodological proxy for validation behaviour.                                                                            |
| Intended interpretation | Ranking or triage score for methodological audit; not autonomous diagnosis, not unknown-pathogen discovery, and not a deterministic Category A lookup replacement.    |
| Threshold policy        | Default $P \geq 0.50$ used for comparability; decision-curve analysis examines threshold dependence.                                                                  |
| Dependence structure    | Country rows nested within reports and reports clustered within disease–time episodes; uncertainty estimated with report grouping and disease–time cluster bootstrap. |

## 2.7 Data-integrity rules and evaluation unit

Dates were parsed using a fixed ISO-compatible conversion. Negative case counts were clipped at zero. Missing case counts were assigned zero in the primary analysis as a plausible default and flagged in the audit; no claim was made that zero represented a confirmed absence of cases. Sensitivity analyses added a missingness indicator and compared zero imputation with training-median imputation plus that indicator. Country names missing in the training file were assigned “Unknown” for GCI construction.

The source data contain multiple country rows for some DON reports. Model fitting occurred at row level, but evaluation occurred at report level using `DONid`. For a multi-country report, the maximum predicted row probability defined the report probability. Cross-validation groups were also defined by `DONid`, preventing rows from the same report from appearing in

248 both training and validation folds.

## 249 **2.8 Leakage-aware feature construction**

250 Two specifications were compared. The spatial–seasonal model (M-SS) used log case count,  
251 month, day of year, the geographic clustering index (GCI), the seasonal mismatch anomaly  
252 (SMA), and the GCI–SMA interaction; the second specification (M-Year) added calendar year  
253 and was otherwise identical.

254 For report row  $i$  with disease  $d$  at time  $t_i$ , GCI was defined as

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

255 where the denominator is the number of unique reporting countries observed on or before  $t_i$ .  
256 Records dated after  $t_i$  are excluded by construction; a symmetric  $\pm 30$ -day window would admit  
257 look-ahead information.

258 SMA was derived from log case counts within each training fold. For disease–month cells  
259 with at least five training observations, the training mean and SD were used. Sparse or absent  
260 disease–month cells used the training month-level reference; a global training reference was  
261 the final fallback. The value was clipped at  $\pm 3$  SD and rescaled to  $[0, 1]$ . Consequently, strict  
262 LOPO did not borrow the held pathogen’s seasonal distribution.

## 263 **2.9 Model specification**

264 An Isolation Forest with 100 trees was fitted to training features, and its negative score was  
265 appended as one additional feature. XGBoost used 200 trees, maximum depth 4, learning rate  
266 0.05, histogram tree construction, and class weighting equal to the training negative-to-positive  
267 ratio. Five random seeds (42, 123, 456, 789, 2024) were prespecified. The reduced ensemble  
268 size was selected for a tractable, fully reproducible audit and was fixed before any results were  
269 computed.

## 2.10 Validation regimes

Naive row-level internal validation used five-fold `StratifiedKFold` on rows, with no group constraint, so that rows from the same report could fall in both the development and the evaluation fold. Grouped internal validation used five-fold `StratifiedGroupKFold` with `DONId` groups. The two designs are otherwise identical—same pipeline, same five seeds, same fold-internal fitting, same aggregation of row probabilities to the report level before scoring—so their difference isolates the grouping choice alone. Temporal holdout trained on 1996–2014 and tested on 2015–2019. Strict LOPO separately removed every row of six agents with at least five rows (Ebola virus, Marburg fever, Lassa fever, plague, anthrax, and tularemia). The 2020–2024 temporal validation trained on the complete 1996–2019 data. Temporal-block internal–external validation used four calendar blocks (1996–2004, 2005–2009, 2010–2014, and 2015–2019), each held out in turn while the remaining periods formed the development data. All seasonal references and model components were fitted using the applicable training subset only.

## 2.11 Cascade scenarios

S0 used individual report predictions. S1 propagated an alert only forward, from a report with  $P \geq 0.50$  to later reports of the same disease within 30 days. S2 additionally normalised three prespecified name pairs: pneumonia of unknown cause to COVID-19, Ebola haemorrhagic fever to Ebola virus, and Marburg haemorrhagic fever to Marburg fever. Cascade performance was evaluated on positives and all negatives.

## 2.12 Second empirical application: hospital readmission

To test the diagnostic on a structurally different clinical task, we used the public Diabetes 130-US-hospitals dataset[46] (101,766 inpatient encounters, 71,518 patients, 1999–2008). The outcome was 30-day readmission (`readmitted = "<30"`); the natural cluster was the patient (`patient_nbr`), with 16,773 patients contributing 47,021 repeat-admission encounters.

Predictors were eight routinely recorded count variables (time in hospital, laboratory procedures, procedures, medications, prior outpatient/emergency/inpatient visits, and number of diagnoses) and thirteen categorical fields (demographics, admission and discharge codes, glucose and HbA1c results, and medication indicators), with “?” treated as missing. A histogram gradient-boosting classifier (200 iterations, maximum depth 6, learning rate 0.06) was evaluated under naive row-level cross-validation and patient-grouped cross-validation (five folds, three seeds each), scored at the encounter level by AUROC. The analysis was run on the full cohort and on the repeat-admission cohort, and the within-patient ICC(1) of both features and outcome was computed as above.

## 2.13 Statistical analysis

Discrimination was quantified by AUC[48, 49]. Threshold performance included sensitivity, specificity, PPV, NPV, and the full confusion matrix. Paired classification differences used a two-sided exact McNemar test[50]. For 2020–2024 uncertainty, reports were grouped into disease-specific clusters separated by more than 90 days; 1,000 cluster bootstrap replicates produced percentile 95% intervals for each model and their paired differences[36]. Confidence intervals were emphasised because validation precision and required sample size depend on event frequency and the target performance measure, not a universal rule of thumb[11, 12]. Calibration was summarised with Brier score, calibration-in-the-large, calibration slope from a logistic calibration model, observed-to-expected ratio, and grouped reliability curves[51, 52]. Decision-curve analysis calculated net benefit across thresholds from 0.05 to 0.60 and was interpreted only as a proxy-outcome stress test[53]. SHAP was not used to support causal interpretation; when feature attribution is reported in future work, it should be treated as a model explanation rather than mechanistic evidence[54].

## 2.14 Reporting and guideline adherence

The study is reported in accordance with TRIPOD+AI[2]; the completed checklist is provided in Additional file 1. Because the object of study is the validation design rather than a model

recommended for use, we also record the PROBAST analysis-domain items relevant to each validation regime we compare[3]. Analyses used Python 3.11 with pandas, scikit-learn, SciPy, Matplotlib, and XGBoost; exact package versions, random seeds, and the released diagnostic utility are provided with the code so that every reported number is reproducible from the raw snapshots.

## 3 Results

### 3.1 Validation optimism in a controlled simulation

We first characterised, under a known data-generating mechanism, how each of three tunable data properties changes the estimate produced by each validation design (Fig. 1; design and approximation in Methods). The approximation predicts two null conditions: row-level leakage should be small if either repeated-cluster information is absent from the predictors or the outcome is not concordant within cluster, and temporal optimism should grow with drift in the prediction mechanism.

Both slices behaved as predicted. When within-cluster correlation was increased while holding the outcome mechanism stationary, report-grouped cross-validation remained stable ( $\text{AUROC} \approx 0.82\text{--}0.85$ ) whereas naive row-level cross-validation inflated monotonically, and the leakage gap widened from 0.8 to 8.9 percentage points as the nuisance correlation grew (Fig. 1a). When concept drift was introduced by rotating the outcome-generating coefficients over time, grouped internal validation became progressively optimistic relative to a strict temporal holdout: the optimism gap widened from 0 to 19.0 percentage points, and under strong drift temporal-holdout discrimination collapsed below 0.5 while internal estimates remained high (Fig. 1b). Varying within-cluster *outcome* concordance while holding feature correlation fixed reproduced the second null condition: the leakage gap was 0.1 percentage points at label  $\text{ICC} \approx 0$  and rose to 7.2 points at label  $\text{ICC} 1.0$ , though the increase was not linear—about half of it was already present by label  $\text{ICC} 0.25$  (Fig. 1c).

The full factorial sweep, however, qualifies the picture in a way the first-order approxima-

tion does not capture, and we report this explicitly because it bounds where the approximation may be used. The optimism gap did behave as an approximately additive function of drift alone: at the highest drift it was 19.2, 15.1, 17.5, 17.9, and 15.9 percentage points across the five feature-ICC levels, with no systematic trend in ICC (Fig. 1e). The leakage gap did not. It grew with feature ICC as expected at zero drift (1.1 to 9.0 percentage points), but it also grew with drift at fixed ICC, more than doubling from 9.0 to 20.8 points at the highest ICC level (Fig. 1d). Most directly, at a measured feature ICC of 0.01—the stated null condition—the leakage gap was 1.1 percentage points under stationarity but 7.0 points under strong drift. The mechanism is straightforward once seen: when the outcome mechanism is non-stationary, cluster membership encodes publication time, and time encodes the locally applicable coefficients, so sharing a cluster between folds leaks the label even when the predictors carry no cluster identity of their own. The product form  $\Delta_{\text{row}} \approx K_f \rho_X \rho_Y$  therefore holds in the stationary regime and understates row-level leakage under drift, and we restrict its use accordingly.

These patterns were robust to model family, with an informative asymmetry (Additional file 1: Table S11). Temporal optimism under drift was essentially model-agnostic: both gradient-boosted trees and logistic regression showed the optimism gap grow to about 20 AU-ROC points at high drift. The row-level leakage gap, however, was flexibility-dependent: it grew with correlation for tree ensembles (to roughly 4–9 points) but stayed near zero for logistic regression, because only a sufficiently flexible learner can memorise cluster identity from nuisance features. Row-level grouping therefore matters most when a flexible model is paired with cluster-correlated data and a concordant outcome. Expressing the axes as empirical ICCs makes the diagnostic usable on real data: an analyst can estimate feature ICC, label ICC, and a drift contrast for a candidate dataset using the released diagnostic utility, and read off the expected order of magnitude. Whether the magnitudes seen here actually arise in real clustered health data is the question the next subsections address.

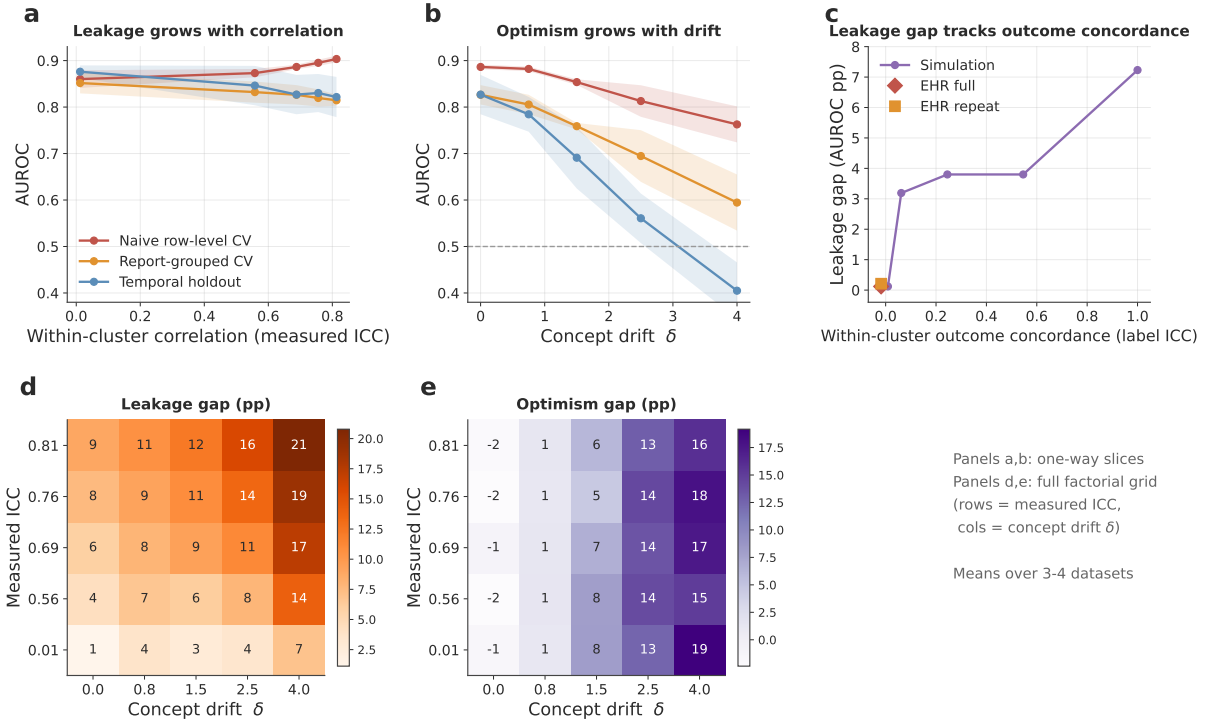


**Fig. 1. Three tunable data properties change what each validation design reports, in a controlled simulation of clustered longitudinal data.** **a**, As within-cluster correlation increases (measured ICC; no drift), report-grouped cross-validation is stable while naive row-level cross-validation inflates. **b**, As concept drift increases (coefficient rotation  $\delta$ ; fixed correlation), grouped internal validation becomes optimistic relative to temporal holdout, which collapses toward and below chance. **c**, The row-level leakage gap (naive – grouped) grows with within-cluster *outcome* concordance (label ICC), though not linearly; the hospital-readmission dataset (markers, outcome ICC  $\approx 0$ ) reproduces the near-zero gap. **d–e**, Full  $5 \times 5$  factorial. The temporal gap (**e**) tracks drift with no systematic trend in ICC. The leakage gap (**d**) does *not* separate: it grows with ICC at zero drift but also with drift at fixed ICC, and reaches 7.0 percentage points in the lowest-ICC row, so the two mechanisms interact and the surfaces should not be read as a look-up table. Lines and markers are means over four generated datasets in **a–c** and three in **d–e**; bands are  $\pm 1$  SD; the dashed line marks chance. Across-repetition standard deviations of individual AUROC estimates range from about 0.5 to 6 percentage points, so small differences between adjacent cells are not interpretable.

### 3.2 Cohort integrity and effective sample size

The 1996–2019 source contained 3,338 country-level rows but only 2,789 unique DON report identifiers. A total of 777 rows belonged to a report identifier appearing more than once. Category A labels occurred in 593 rows but 459 unique reports. Case counts were missing in 1,788 rows (53.6%). The 2020–2024 dataset contained 301 rows, 298 unique reports, 74 Category A rows, and 73 Category A reports; 172 rows (57.1%) had missing case counts. Missing case counts were retained with a zero value in the primary analysis to mirror a common default,

but this was treated as an assumption and examined in missingness-aware sensitivity analyses (Additional file 1: Table S7).

All primary performance results below use the DON report as the evaluation unit. For multi-country reports, the maximum row-level probability was used because the operational question is whether any component of a report triggers review. This reduced pseudoreplication but does not remove correlation between successive reports from the same outbreak; cluster bootstrap intervals were therefore used for the later-period validation.

### **3.3 Grouped cross-validation substantially overestimated temporal performance**

Removing the grouping constraint changed the internal estimate very little. Naive row-level five-fold cross-validation gave AUC  $0.942 \pm 0.001$  for M-SS and  $0.968 \pm 0.002$  for M-Year, against  $0.935 \pm 0.003$  and  $0.962 \pm 0.001$  for the report-grouped design—leakage gaps of 0.76 and 0.49 percentage points within the matched diagnostic run (Table 2, footnote b). This is an order of magnitude smaller than the simulation’s high-ICC cells would suggest, and we return to why below. In the 2015–2019 temporal holdout, AUC fell to  $0.617 \pm 0.008$  for M-SS and  $0.804 \pm 0.007$  for M-Year (Table 2; Fig. 2; per-seed values in Additional file 1: Tables S2 and S3). The corresponding CV-to-temporal gaps were 31.8 and 15.8 percentage points. An M-SS temporal-holdout AUC of 0.617 is close to chance (0.5); the honest reading is therefore a negative result—the spatial–seasonal model did not transport to the held-out period—rather than evidence of moderate retained performance. Under leakage-aware validation, calendar year was therefore not the primary driver of the temporal performance gap. Year retained predictive information in the 2015–2019 holdout, while the year-free model showed greater optimism.

**Table 2.** Leakage-aware validation performance at report level

| Validation                             | Model  | AUC         | Sensitivity | Specificity | PPV                |
|----------------------------------------|--------|-------------|-------------|-------------|--------------------|
| Naive row-level 5-fold CV <sup>b</sup> | M-SS   | 0.942±0.001 | 0.902       | 0.838       | —                  |
|                                        | M-Year | 0.968±0.002 | 0.913       | 0.893       | —                  |
| Grouped 5-fold CV                      | M-SS   | 0.935±0.003 | 0.899       | 0.828       | —                  |
|                                        | M-Year | 0.962±0.001 | 0.902       | 0.885       | —                  |
| Temporal holdout, 2015–2019            | M-SS   | 0.617±0.008 | 0.492       | 0.853       | 0.461 <sup>a</sup> |
|                                        | M-Year | 0.804±0.007 | 0.593       | 0.780       | 0.405 <sup>a</sup> |
| Temporal validation, 2020–2024         | M-SS   | 0.831       | 0.795       | 0.640       | 0.417              |
|                                        | M-Year | 0.797       | 0.740       | 0.671       | 0.422              |

PPV, positive predictive value. Threshold metrics use  $P \geq 0.50$ . Internal-validation and temporal-holdout values are means across five seeds; AUC includes SD. The two temporal rows use different development windows (1996–2014 and 1996–2019 respectively) and are therefore not a monotone ladder in forecast horizon. <sup>a</sup>Mean of seed-specific PPVs. <sup>b</sup>Naive row-level cross-validation was computed with the identical pipeline, seeds, and report-level aggregation as the grouped design, differing only in that folds were not constrained by `DONid`. In that same run, grouped cross-validation returned 0.9343 (M-SS) and 0.9633 (M-Year), reproducing the tabulated values to within 0.15 percentage points; the within-run leakage gaps were therefore 0.76 and 0.49 percentage points.

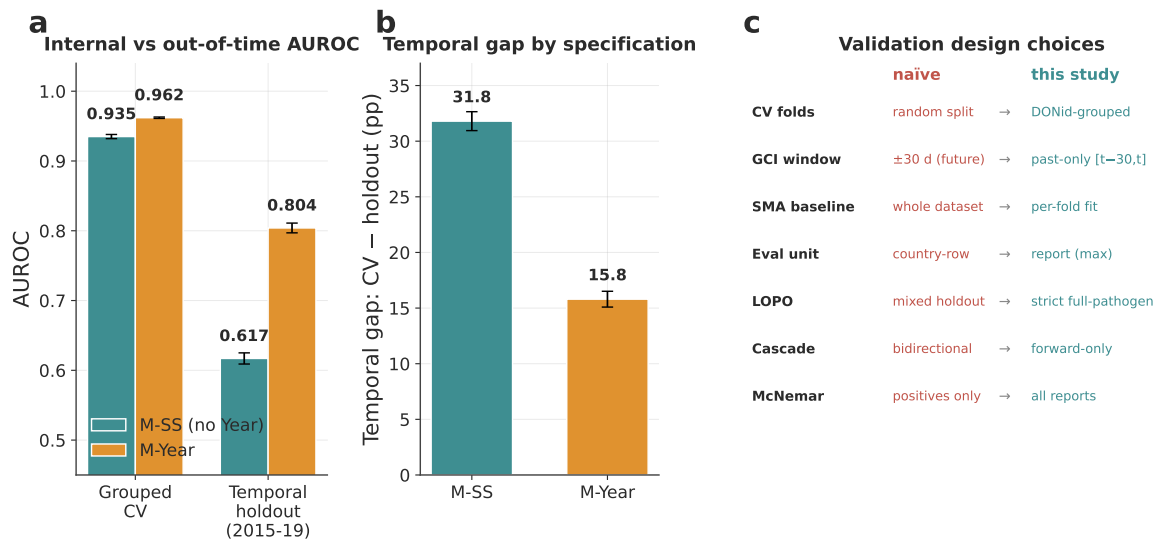


**Fig. 2. Leakage-aware validation changes apparent performance.** **a**, Report-grouped five-fold cross-validation versus 2015–2019 temporal-holdout AUROC for M-SS and M-Year (mean±SD across five seeds). **b**, The temporal gap  $\Delta_{\text{time}}$  (grouped-CV minus temporal-holdout AUROC) by model. **c**, The seven design choices that distinguish a conventional row-level analysis of these data from the leakage-aware one.

### 3.4 Strict leave-one-pathogen-out performance was modest

Strict leave-one-pathogen-out (LOPO) removed all rows of each held pathogen from training and fitted the seasonal reference only on the remaining data. Report-level detection varied substantially (Fig. 3). M-SS detection ranged from 28.9% for Anthrax to 72.2% for Plague; M-Year ranged from 28.0% for Tularemia to 73.3% for Anthrax. Macro-average detection

was 50.0% for M-SS and 47.7% for M-Year (per-pathogen values, mean $\pm$ SD across seeds, in Additional file 1: Table S4). A non-strict holdout that mixes an epidemic-level Ebola exclusion with pathogen-level exclusions for the other agents gives 27.3% instead; that quantity is not a cross-pathogen estimand and the two should not be compared.

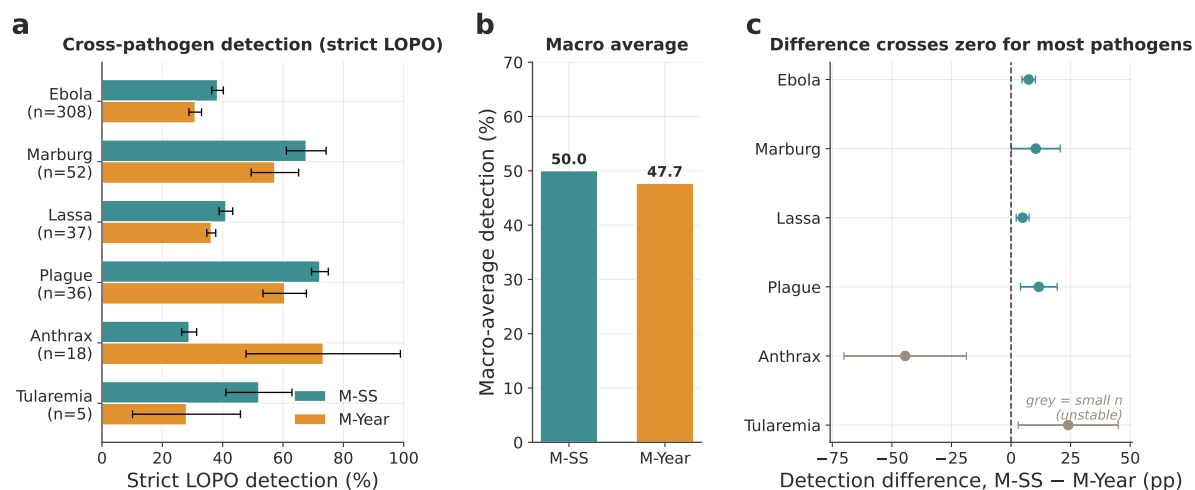


**Fig. 3. Strict leave-one-pathogen-out (LOPO) validation.** **a**, Report-level detection for each held-out pathogen (mean $\pm$ SD across five seeds;  $n$  = held-out reports); every row of the named pathogen was removed from both training and the seasonal reference. **b**, Unweighted macro-average detection across the six prespecified pathogens. **c**, Per-pathogen difference (M-SS minus M-Year) with combined SD; differences scatter around zero, and small- $n$  pathogens (grey) are unstable. LOPO estimates are descriptive and threshold-dependent.

### 3.5 Later-period validation showed no statistically significant model superiority

The 2020–2024 later-period validation contained 298 reports, including 73 Category A reports. M-SS produced AUC 0.831, sensitivity 79.5% (58/73), specificity 64.0% (144/225), PPV 41.7% (58/139), and Brier score 0.196. M-Year produced AUC 0.797, sensitivity 74.0% (54/73), specificity 67.1% (151/225), PPV 42.2% (54/128), and Brier score 0.197 (Fig. 4; report-level confusion matrices and cluster-bootstrap intervals in Additional file 1: Table S5).

Cluster bootstrap used 137 disease–time clusters. M-SS AUC had a 95% interval of 0.741–0.884; the AUC difference versus M-Year was 0.034 (95% interval –0.014 to 0.101). The sensitivity difference was 5.5 percentage points (95% interval –3.0 to 18.2), specificity difference –3.1 points (95% interval –9.6 to 2.5), and PPV difference –0.5 points (95% interval

422 –5.6 to 4.6). On Category A reports, discordant predictions were 7 M-SS-only detections ver-  
 423 sus 3 M-Year-only detections (two-sided exact McNemar  $p = 0.344$ ). Across all reports, the  
 424 numbers exclusively classified correctly were 24 and 27 ( $p = 0.780$ ). The data therefore did  
 425 not demonstrate threshold-level superiority of either model.

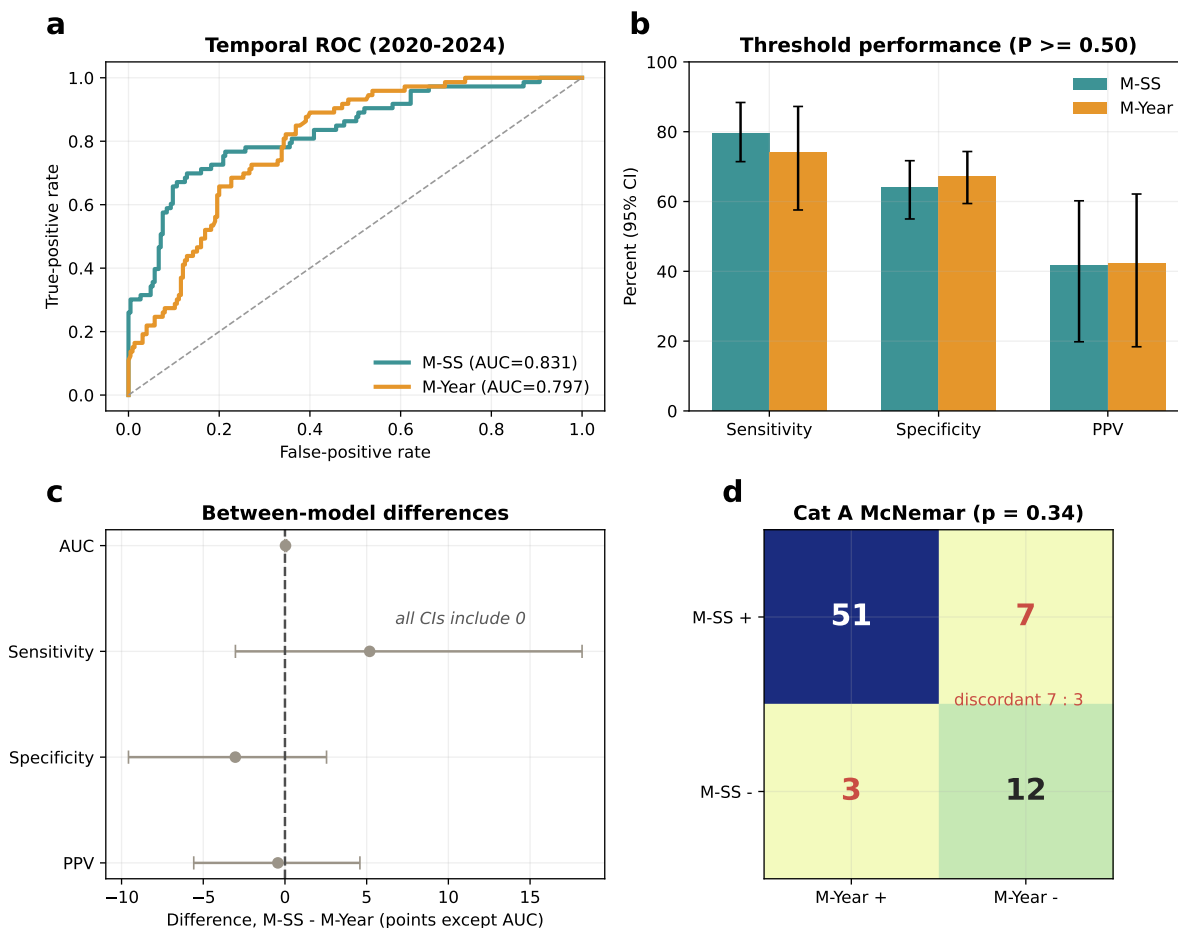


**Fig. 4. Later-period validation (2020–2024; 298 reports, 73 Category A; same WHO source, not a prospective cohort).** **a**, Receiver operating characteristic curves (AUC 0.831 for M-SS, 0.797 for M-Year). **b**, Sensitivity, specificity, and PPV at  $P \geq 0.50$  with cluster-bootstrap 95% intervals. **c**, Between-model differences (M-SS minus M-Year) with 95% intervals; all four include zero. **d**, Paired McNemar contingency for Category A detection (discordant 7 vs 3; two-sided exact  $p = 0.344$ ).

### 426 3.6 Cascade propagation increased sensitivity and false-alert burden

427 Without cascade propagation (S0), M-SS detected 58/73 positive reports and generated 81  
 428 false positives. A forward-only 30-day cascade within disease clusters (S1) detected 69/73  
 429 positive reports but generated 93 false positives, reducing specificity from 64.0% to 58.7%.

PPV changed only from 41.7% to 42.6%. Adding the prespecified nomenclature map (S2) did not change aggregate performance in this report-level analysis. The cascade is therefore a sensitivity–specificity tradeoff, not an unqualified improvement (Additional file 1: Table S10; Additional file 1: Fig. S1).

### **3.7 Calibration, missingness, decision-curve, and temporal-block stress tests**

Later-period probabilities were not well calibrated for direct resource allocation. Mean predicted risk exceeded the observed Category A report proportion for both models (M-SS 46.4% predicted versus 24.5% observed; M-Year 42.5% versus 24.5%), with observed-to-expected ratios of 0.53 and 0.58, respectively. Calibration slopes were 1.08 for M-SS and 0.70 for M-Year, indicating that calibration differed by specification even when Brier scores were similar (Fig. 5a; Additional file 1: Table S6).

A decision-curve illustration supported threshold-dependent rather than universal usefulness. For M-SS, net benefit was positive at lower risk thresholds (0.167 at threshold 0.10 and 0.109 at threshold 0.20), but became negative at threshold 0.50 (−0.077), consistent with the low PPV and high false-alert burden at the default classification threshold (Fig. 5b; Additional file 1: Table S9). Because the outcome is a methodological proxy, these net-benefit values should be read as an interpretive stress test rather than a deployment utility claim.

Case-count missingness did not explain away the main temporal-validation result, but it mattered. Adding a missingness indicator to the zero-imputation model slightly increased AUC from 0.831 to 0.836, sensitivity from 79.5% to 82.2%, specificity from 64.0% to 65.3%, and PPV from 41.7% to 43.5%. Replacing missing counts with the training median plus a missingness indicator reduced AUC to 0.748 and PPV to 36.5% (Fig. 5c; Additional file 1: Table S7). Thus, missingness handling is a material modelling choice and should be reported as part of validation, not hidden inside preprocessing.

Temporal-block internal–external validation across four calendar blocks showed large non-stationarity. AUCs were high for 1996–2004 and 2005–2009, poor for 2010–2014 (0.464 for

457 M-SS and 0.479 for M-Year), and divergent in 2015–2019 (0.617 versus 0.804). This hetero-  
 458 geneity reinforces that a single random internal-validation summary is insufficient for clustered  
 459 longitudinal reporting systems (Fig. 5d; Additional file 1: Table S8).

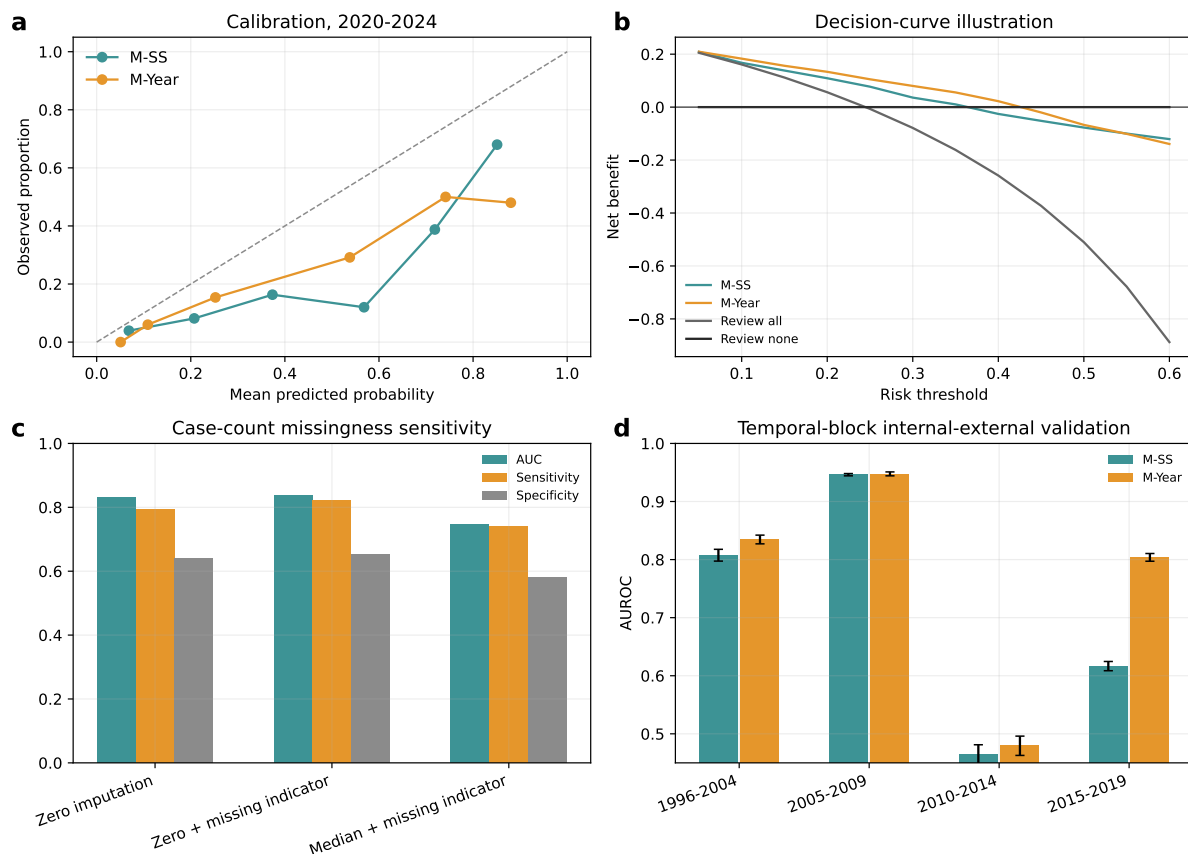


**Fig. 5. Methodological stress tests requested by prediction-model validation guidance.** **a**, Calibration curves for the 2020–2024 later-period validation set. **b**, Decision-curve illustration for proxy triage thresholds; net benefit is threshold-dependent and should not be interpreted as operational utility without an explicit cost model. **c**, Case-count missingness sensitivity for M-SS in 2020–2024. **d**, Temporal-block internal–external validation across four calendar periods (mean±SD across five seeds).

### 460 3.8 Transportability to a genuinely predictive outcome

461 Because Category A status is deterministic given the disease name, we repeated the validation  
 462 ladder on a second, non-deterministic outcome to test whether the same optimism pattern arises  
 463 for a genuine forecasting target. We defined *report escalation*: for each report, whether the  
 464 same disease produced at least one further WHO DON report within the subsequent 90 days  
 465 (prevalence 86.0%,  $n = 2,755$  reports with complete follow-up); the outcome uses only future

reporting for its definition while features remain past-only. Using the identical M-SS feature set, naive row-level cross-validation gave AUC 0.823, report-grouped cross-validation gave 0.809 (a 1.4-percentage-point leakage gap), and 2015–2019 temporal holdout gave 0.751 (a further 5.8-percentage-point optimism gap; Additional file 1: Fig. S2). The same ordering—naive > grouped > temporal—reproduced on this genuinely predictive outcome, indicating that the leakage-aware recommendations are not artefacts of the deterministic proxy but transfer to a real forecasting task on the same archive.

### 3.9 A second clinical dataset refines when row-level grouping matters

To probe transportability beyond surveillance data and to test the limits of the diagnostic, we applied it to a structurally different clinical prediction task: 30-day hospital readmission in the public Diabetes 130-US-hospitals dataset (101,766 encounters from 71,518 patients; 47,021 encounters from the 16,773 patients with repeat admissions; Methods). Here the natural cluster is the patient rather than the report. As in the surveillance archive, naive row-level and patient-grouped cross-validation gave almost identical discrimination—AUROC 0.6737 versus 0.6725 in the full cohort ( $\Delta_{\text{row}} = 0.12$  percentage points) and 0.6215 versus 0.6194 among repeat-admission patients ( $\Delta_{\text{row}} = 0.21$  points)—even though the routine clinical features were moderately correlated within patient (mean feature ICC 0.37).

This near-zero gap is predicted by the diagnostic and sharpens it. Row-level leakage requires that cluster membership carry information about the *label*, not merely about the features: a model can only exploit a shared training/test cluster if knowing the cluster narrows the outcome. In the readmission task the within-patient outcome concordance was essentially nil (label ICC  $-0.02$ )—whether a given admission is followed by readmission is largely independent across a patient’s admissions—so memorising patient identity does not leak the label, and grouping is nearly irrelevant. A dedicated simulation that varied outcome concordance while holding feature correlation fixed confirmed the mechanism: the leakage gap grew from 0.1 to 7.2 AUROC points as the within-cluster label ICC increased from 0 to 1, and the zero-concordance end of the simulation coincided with the observed readmission gap (Fig. 1c).

Within-cluster outcome concordance is therefore a necessary condition for row-level leakage, and the label ICC identifies its absence before a model is trained. It is not a sufficient condition, as the surveillance archive shows: the next subsection places both datasets on all of the relevant axes.

### 3.10 Diagnostic coordinates of the two empirical datasets

Table 3 places both datasets on the axes the simulation varied. The two occupy opposite corners of the ICC space, and the surveillance archive sits at the theoretical maximum of the leakage condition: the Category A label is a deterministic function of the reported disease, which is shared by every row of a report, so the within-report label ICC is exactly 1.00, and the persistent predictors have a mean within-report ICC of 0.83. On the ICC criterion alone this is the most leakage-prone dataset the approximation admits. The measured leakage gap was nevertheless 0.76 percentage points.

The reconciling quantity is cluster size, which ICC does not encode. A high label ICC states that knowing the cluster determines the outcome; it says nothing about how often a learner will actually see the same cluster on both sides of a fold. In the surveillance archive the 3,338 rows fall into 2,789 reports, a mean of 1.20 rows per report, and only 228 reports (8.2%) contribute more than one row—so for 92% of the evaluation units there is no within-cluster information to leak, whatever the ICC. The readmission cohort is more repeated in this sense (1.42 encounters per patient overall, 23.5% of patients with repeat admissions) but has essentially no outcome concordance to leak. Both datasets are therefore leak-resistant, for opposite reasons, and the simulation reaches its large leakage gaps only because its clusters average 2.5 rows with high concordance throughout. Row-level leakage requires the conjunction of all three conditions, and in these two real datasets the conjunction did not hold.

The temporal axis behaved differently. The drift contrast between the 1996–2014 development period and the 2015–2019 holdout was  $D = 0.37$ , and the corresponding gap between grouped internal and temporal-holdout AUC was 31.8 percentage points for M-SS—roughly forty times larger than anything the grouping choice produced in the same data. Two obser-

variations qualify a simple drift reading, however, and we report them because they bound the temporal approximation as the factorial sweep bounded the leakage one. First, the contrast between 1996–2019 and 2020–2024 was larger still ( $D = 1.00$ , near-orthogonal coefficient vectors) yet gave a *higher* AUC (0.831) than the nearer 2015–2019 holdout (0.617); the two are not comparable because the later evaluation was trained on five additional years of data, so temporal degradation in this archive is not monotone in forecast horizon. Second, the 2005–2009 block returned an AUC of 0.946 for M-SS, marginally above grouped cross-validation—negative optimism. What the archive shows is therefore strong between-period heterogeneity rather than uniform, horizon-proportional drift, and  $D$  ranks the periods correctly only when the development window is held fixed.

**Table 3.** Pre-modelling diagnostic coordinates and observed validation gaps for the two empirical datasets

| Quantity                                                    | WHO DON archive       | Hospital readmission        |
|-------------------------------------------------------------|-----------------------|-----------------------------|
| Natural cluster                                             | DON report            | Patient                     |
| Rows (evaluation units)                                     | 3,338 (2,789 reports) | 101,766 (71,518 patients)   |
| Mean rows per cluster                                       | 1.20                  | 1.42                        |
| Clusters with $>1$ row                                      | 8.2%                  | 23.5%                       |
| Label ICC(1) within cluster                                 | 1.00                  | −0.02                       |
| Mean feature ICC(1) within cluster <sup>a</sup>             | 0.83                  | 0.37                        |
| Drift contrast $D$ (development $\rightarrow$ holdout)      | 0.37                  | not applicable <sup>b</sup> |
| $\Delta_{\text{row}}$ : naive – grouped CV (AUC, pp)        | 0.76                  | 0.12 (0.21) <sup>c</sup>    |
| $\Delta_{\text{time}}$ : grouped CV – temporal holdout (pp) | 31.8                  | not evaluated               |

ICC, intraclass correlation coefficient; pp, percentage points. <sup>a</sup>In the surveillance archive the seasonal mismatch anomaly and its interaction term are fitted inside training folds and have no pre-modelling column; the mean is over the four persistent predictors. For the readmission cohort the feature ICC is computed on the repeat-admission subcohort over the eight numeric predictors, since singleton clusters carry no within-cluster information. <sup>b</sup>The readmission cohort was analysed as a cross-sectional cluster problem; no temporal split was defined. <sup>c</sup>Full cohort (repeat-admission cohort).

## 4 Discussion

### 4.1 Principal findings

The simulation and the two empirical datasets answer the question in the title differently, and the difference is the main result. Under a known data-generating mechanism, validation design

534 mattered a great deal and in the way the approximation predicts: within-cluster feature corre-  
535 lation and outcome concordance acted jointly on row-level leakage, so that either alone left the  
536 naive and grouped estimates in agreement, and drift acted on the gap between grouped inter-  
537 nal and temporally separated estimates. In the two real datasets, the grouping choice barely  
538 mattered at all—0.76 and 0.12 percentage points of AUC—while the temporal contrast in the  
539 archive that permitted one was 31.8 points.

540 The reconciliation is instructive rather than deflationary. Intraclass correlation states how  
541 strongly cluster membership determines a value; it is silent on how often membership is actually  
542 shared. The surveillance archive attains the theoretical maximum of the ICC condition—a label  
543 that is a deterministic function of a report-level attribute, hence label ICC 1.00—and still leaks  
544 almost nothing, because 92% of its reports contribute a single row. The readmission cohort  
545 has more repetition but no outcome concordance to leak. Row-level leakage therefore requires  
546 three conditions in conjunction, not two: cluster-identifying predictors, within-cluster outcome  
547 concordance, and enough repeated observations per cluster for a learner to exploit. The third  
548 is the one usually left unstated, and in both of our datasets it was the binding one. An analyst  
549 who checks only ICC will over-predict the benefit of grouped resampling; an analyst who  
550 checks cluster size as well will often find, as we did, that the constraint can be relaxed without  
551 consequence.

552 We also found the factorial sweep to be less tidy than the first-order form implies. The  
553 leakage gap grew not only with feature ICC but also with drift at fixed ICC, more than dou-  
554 bling at the highest ICC level, and reached 7.0 percentage points at a measured feature ICC  
555 of 0.01 where the approximation predicts approximately none. Under non-stationarity, cluster  
556 membership proxies publication time and time proxies the locally applicable coefficients, so  
557 the two mechanisms are not separable. The product form is a stationary-regime approximation,  
558 and we have marked it as such rather than presenting the grid as a look-up table.

559 The surveillance application shows what a consequential validation contrast looks like when  
560 there is one. Once seasonal references were fitted inside the training folds, the clustering index  
561 was restricted to already-available information, strict pathogen holdouts were used, and cali-  
562 bration, missingness, and decision-analytic stress tests were added, the conclusion changed in

kind and not only in degree: internal validation reported AUCs of 0.935 and 0.962, whereas temporally separated validation returned 0.617 and 0.804, and no between-model difference survived cluster-aware uncertainty. M-SS showed a numerically higher AUC and sensitivity than M-Year in 2020–2024, but the uncertainty intervals included no difference and PPV remained approximately 42% for both models.

The grouped-CV results remain optimistic. Even after grouping by report identifier, neighbouring DON reports may describe the same outbreak, and temporal holdout performance was substantially lower. This agrees with wider evidence that leakage, clustered observations, and acquisition shifts can make internal ML validation unreliable[24, 25, 34]. The temporal-block internal–external analysis confirmed substantial heterogeneity between calendar periods, including a low-performing 2010–2014 block. Internal–external cross-validation, rolling-origin evaluation, or explicit cluster-level validation should therefore be preferred when multiple outbreak clusters, sites, health systems, or calendar periods are available[13].

Year should not be described simply as a nuisance confounder. In the 2015–2019 holdout, M-Year substantially outperformed M-SS, suggesting that calendar time carried transportable information over that interval. In 2020–2024, M-SS ranked reports somewhat better, but the difference was uncertain. The practical lesson is to evaluate temporal features under multiple forward windows rather than exclude them solely because they correlate with historical clusters.

The strict LOPO analysis also changes the interpretation of cross-pathogen transfer. Strict removal of an entire pathogen gave about 50% macro detection for both specifications, with strong heterogeneity by agent. This is not evidence that the system can discover a genuinely novel disease. Both GCI and SMA require a disease grouping, and Category A status can be obtained directly once a disease is named. The benchmark is therefore useful for studying validation behaviour, not for claiming automated pathogen identification.

Complete error accounting is essential. A positives-only reading would omit the 225 negative reports and obscure the alert burden. Likewise, cascade propagation raised sensitivity but lowered specificity. Whether such a tradeoff is acceptable depends on review capacity and the relative cost of missed versus unnecessary alerts. The decision-curve illustration made this de-

pendence visible: the same model could have positive net benefit at lower triage thresholds and negative net benefit at the default 0.50 threshold. Formal deployment decisions would require stakeholder-defined thresholds and a cost model rather than a universal claim of benefit[53].

Probability estimates were imperfect: Brier scores were about 0.196, mean predicted risks exceeded observed prevalence, and threshold performance was sensitive to specification. Calibration should be reported with uncertainty and evaluated before probabilities are used for resource allocation[51, 52]. In this analysis, probabilities are treated as ranking and triage scores only.

## 4.2 Comparison with existing literature

The closest precedent is Wynants et al., who showed in a simulation of multicentre data that ignoring clustering alters calibration and discrimination, with the intraclass correlation as the governing quantity and a heavily clustered condition (ICC around 20%) as the regime of concern[43]. Our contribution is complementary in three respects. We separate the feature-side from the outcome-side component of clustering and show that both are required; we add the temporal axis and show that in a real archive it dominates the clustering axis by roughly forty-fold; and we carry the diagnostic through to two empirical datasets, where the leakage magnitudes the simulation implies did not materialise. Little, Varoquaux and colleagues, in a joint commentary with the original authors, had already argued that the appropriate cross-validation scheme depends on the inference intended rather than being fixed in advance[44]; the present study supplies the quantities that make that judgement operational, and—unexpectedly—finds the permissive side of their argument to be the empirically common case.

Our results are also consistent with reports that single-study external validation gives highly variable performance estimates depending on which cluster is used, and that internal–external cross-validation is better equipped to expose that variability[13, 11], and with the wider observation that high internal machine-learning performance frequently fails to survive shifts in site, acquisition, or calendar period[24, 25, 34]. The temporal-block analysis reported here, with a low-performing 2010–2014 block, a divergent 2015–2019 block, and a 2005–2009 block that

slightly exceeded internal validation, is a concrete instance of the between-cluster heterogeneity that motivates internal–external designs—and a caution against reading such heterogeneity as monotone temporal decay. For studies of this kind, TRIPOD-Cluster provides reporting recommendations specific to clustered and heterogeneous data that complement TRIPOD+AI[55, 56]; we have taken its clustering-specific recommendations into account throughout.

Where we depart from much of this literature is in the sign of the practical conclusion. The prevailing message is that analysts under-protect against clustering. On the two datasets examined here the opposite held: the clustering protection was nearly costless to omit, and the under-protected axis was time.

These results should not be read as scepticism about prediction in non-stationary settings in general. What we document is a property of how a model is validated on clustered, drifting data, not an intrinsic ceiling on what can be predicted from such data. The distinction the framework enforces—between discrimination that survives grouped and temporally separated evaluation and discrimination that does not—is what separates a transportable model from an artefact of the validation design. Reporting both the optimistic and the honest estimate, as we do here, makes that distinction auditable rather than rhetorical.

### 4.3 Strengths and limitations

The main strength of this study is that the same validation ladder was applied across a controlled simulation and two empirical datasets with deliberately opposite cluster structures, so the mechanism is identified under a known data-generating process and then observed at both of the boundary conditions it predicts. Because the diagnostic quantities are ordinary intraclass correlation coefficients and a drift contrast, they can be computed from any dataset with a cluster identifier and a time stamp, without access to a fitted model. All raw snapshots, analysis code, seeds, and the diagnostic utility are released, so every reported number can be recomputed.

Five limitations qualify these conclusions. Before them, one caveat applies to the simulation specifically: cells of the factorial sweep were averaged over three generated datasets and

the one-dimensional slices over four, which is adequate for the order-of-magnitude comparisons we draw but not for resolving small differences between adjacent cells. Across-repetition standard deviations of the individual AUROC estimates ranged from about 0.5 to 6 percentage points depending on the cell, so with three or four repetitions the Monte Carlo error on a single cell is of the order of a percentage point or more, and the non-monotone cells visible in the lowest-ICC row of Fig. 1d should not be read as structure. A sweep with several hundred repetitions per cell, reported with Monte Carlo standard errors throughout as recommended for simulation studies[47], would be required to characterise the surfaces precisely. First, more than half of rows lacked a case count; sensitivity analyses showed that missingness handling changed performance, so zero imputation should not be treated as an epidemiological truth, and more principled approaches such as multiple imputation warrant evaluation in future work[57, 58, 59]. Second, the 2020–2024 dataset came from the same WHO reporting system and is a later-period validation, not independent prospective collection. Third, the primary outcome is a deterministic proxy used as a methodological test-bed; a secondary genuinely-predictive outcome (report escalation) reproduced the same validation pattern and the controlled simulation established the governing factors, but broader replication on independent real-world datasets remains future work. Fourth, report grouping does not eliminate within-outbreak dependence; cluster bootstrap and temporal-block validation only partially address this structure. Reports were aggregated by the maximum row-level probability, but mean aggregation changed discrimination negligibly (AUROC differences of at most about 0.5 percentage points for both grouped cross-validation and later-period validation, and only 228 of 2,789 reports contain multiple rows; Additional file 1), so the aggregation choice does not affect the conclusions. Fifth, model hyperparameters were fixed rather than tuned, and a threshold of 0.50 was not selected through a prospective utility analysis.

One practical property of the diagnostic is worth stating. It requires only a cluster identifier, a time stamp, the predictors, and the outcome—no fitted model, no specialised software, and no access to the development pipeline—so it can be computed by a reviewer or by an analyst working without statistical infrastructure.

## 4.4 Recommendations

The broader implication is methodological and, we argue, general rather than archive-specific. It also runs partly against the direction in which the field’s advice currently points. The established recommendation to use subject- or record-aware cross-validation is stated without qualification[18, 17]; our results refine it in both directions. Grouped resampling is necessary when cluster-identifying predictors, within-cluster outcome concordance, and substantial repetition per cluster occur together—and when they do, the simulation shows the cost of omitting it can exceed 20 AUROC points. But that conjunction was absent from both real datasets we examined, and enforcing grouping there bought between 0.1 and 0.8 points at the price of a smaller effective sample. Meanwhile the temporal contrast, which no comparably prominent recommendation makes mandatory, moved the archive’s headline result by 31.8 points. On the evidence here, an analyst working with clustered longitudinal health data should spend the marginal effort on temporal validation before spending it on grouping, and should decide the grouping question from the three measured quantities rather than by default. The same reasoning applies to electronic health records with repeated encounters[60, 61, 62], disease registries, and wearable time series, not only to outbreak reports. We distil the audit into seven leakage-aware validation recommendations; each restates an established principle, and our contribution is to quantify the magnitude of its effect and the data conditions under which it matters (Table 4):

1. Define the real evaluation unit (e.g., the report or the patient, not the row) and group all resampling by it.
2. Fit every data-derived transformation—baselines, normalisation, imputation, encoders—inside the training folds only.
3. Prohibit future information from entering any feature intended for real-time use.
4. Apply identical, strict holdout definitions across subgroups whenever cross-group transfer is claimed.
5. Restrict post-processing rules (e.g., alert cascades) to information available at decision

time, and report their cost as well as their benefit.

6. Report the full confusion matrix—sensitivity, specificity, PPV/NPV, calibration, and decision-analytic consequences—rather than positives-only agreement.

7. Quantify uncertainty with resampling that respects the clustering (e.g., cluster bootstrap), and interpret differences against confidence intervals rather than point estimates.

Reporting should follow current TRIPOD+AI and risk-of-bias guidance where applicable[2, 3, 4]. Together these practices provide a more defensible basis for deciding whether a model merits operational evaluation[63, 64, 65].

## 5 Conclusions

For clustered longitudinal health data, validation design is part of the scientific claim rather than a downstream technical detail—but which part of the design carries the weight is a property of the data, not a constant. Row-level leakage requires three conditions in conjunction: predictors that carry cluster identity, an outcome that is concordant within clusters, and enough repeated observations per cluster for a flexible learner to exploit. All three can be measured before a model is fitted. In the two real datasets examined here the conjunction failed—in one because clusters were almost all singletons despite a label ICC of 1.00, in the other because the outcome was not concordant within patients—and grouped resampling changed discrimination by less than one percentage point. Temporal separation, by contrast, moved the surveillance archive’s headline result by 31.8 percentage points and reversed its conclusion: grouped internal validation remained optimistic, temporal performance was heterogeneous rather than monotonically decaying, model differences in 2020–2024 were uncertain, calibration was inadequate for direct resource allocation, and cascade processing traded missed alerts for false alerts.

The practical recommendation follows from the asymmetry. Report the three clustering diagnostics and let them decide whether grouped resampling is needed, rather than applying it by default at a cost in effective sample size; and treat temporally separated validation as the contrast that must be reported, not the optional extra. Beyond that, the audit we used remains:

**Table 4.** Leakage-aware recommendations, their established basis, and quantified effect

| Recommendation                                                                     | Established in                                   | Quantified effect in this study                                                                                                                                                                                                                                           |
|------------------------------------------------------------------------------------|--------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Group resampling by the evaluation unit                                            | TRIPOD+AI; Riley IECV; TRIPOD-Cluster[2, 13, 55] | Grouping by report changed internal AUC by only 0.76 percentage points here (0.942 to 0.934 for M-SS within the matched run), because 92% of reports contribute one row; the same choice moved the simulation by up to 20.8 points where clusters were larger.            |
| Fit data-derived transformations inside training folds                             | PROBAST analysis domain[3]                       | Fitting seasonal baselines per fold rather than on all data was required to avoid inflated temporal performance.                                                                                                                                                          |
| Block future information from real-time features                                   | TRIPOD+AI[2]                                     | Restricting the clustering window to $[t - 30, t]$ removed look-ahead information present in a symmetric $\pm 30$ -day window.                                                                                                                                            |
| Apply identical strict holdouts across subgroups                                   | Rarely formalised                                | Strict leave-one-pathogen-out gave $\sim 50\%$ macro detection, versus 27.3% under a non-strict holdout mixing epidemic- and pathogen-level exclusions.                                                                                                                   |
| Restrict post-processing to decision-time information and report its cost          | Decision-curve analysis[53]                      | A forward-only cascade raised sensitivity but lowered specificity (false positives 81 $\rightarrow$ 93); at the default 0.50 threshold the underlying model already had negative net benefit ( $-0.077$ ), so the cascade added alerts where alerting was not beneficial. |
| Report the full confusion matrix and calibration                                   | TRIPOD+AI; calibration hierarchy[2, 51]          | All-report accounting and calibration (slope 1.08 and 0.70; observed-to-expected 0.53 and 0.58) qualified a positives-only, discrimination-only reading.                                                                                                                  |
| Quantify uncertainty with cluster-aware resampling and interpret against intervals | Sample-size and IECV guidance[12, 13]            | The cluster-bootstrap AUC-difference interval ( $-0.014$ to $0.101$ ) crossed zero, so the numerically higher M-SS AUC could not be read as a model advantage.                                                                                                            |

Each recommendation restates an established validation principle; the contribution of this study is the quantified effect of the corresponding source on conclusions drawn from clustered longitudinal health data.

726 define the evaluation unit, fit transformations only inside training data, prohibit future infor-  
727 mation, use strict holdouts for transfer claims, report complete threshold errors and calibration,  
728 quantify cluster-aware uncertainty, and interpret the model only within the intended-use setting.  
729 A model that survives that ladder may merit operational evaluation; one that does not should  
730 be treated as an internal signal rather than a validated prediction model.

## Abbreviations

AUC/AUROC: area under the receiver operating characteristic curve; CDC: United States Centers for Disease Control and Prevention; CV: cross-validation; DON: Disease Outbreak News; GCI: geographic clustering index; ICC: intraclass correlation coefficient; LOPO: leave-one-pathogen-out; M-SS: spatial–seasonal model specification; M-Year: spatial–seasonal model specification with calendar year; NPV: negative predictive value; PPV: positive predictive value; PROBAST: Prediction model Risk Of Bias ASsessment Tool; SMA: seasonal mismatch anomaly; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; WHO: World Health Organization.

## Declarations

### Ethics approval and consent to participate

Not applicable. The study used publicly available, aggregate outbreak reports and a publicly available, de-identified hospital dataset. It did not involve identifiable individual-level data, participant contact, or intervention, so institutional review board approval and informed consent were not required.

### Consent for publication

Not applicable.

### Availability of data and materials

All data and code supporting the conclusions of this article are available in the antipandemic repository, <https://github.com/METX-915/antipandemic>. The repository contains the raw 1996–2019 and 2020–2024 snapshots used here, report-level predictions, data-integrity summaries, bootstrap outputs, the simulation and hospital-readmission analy-

ses, all figure scripts, and the released leakage diagnostic (`leakage_diagnostic.py`). The 1996–2019 source dataset is additionally available from the CGHSS/Georgetown DON repository[45] and the 2020–2024 source reports through the WHO DON API; reuse must preserve WHO source attribution. The hospital-readmission data are the public Diabetes 130-US-hospitals dataset[46].

## Competing interests

The authors declare that they have no competing interests.

## Funding

This research was conducted as part of a project under the Ministry of National Defense of the Republic of Korea. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; or in writing the manuscript.

## Authors' contributions

HN: conceptualisation, methodology, software, investigation, writing — original draft. JY: data curation, investigation, software, writing — original draft. DHL: data curation, formal analysis, visualisation, writing — original draft. JK: methodology, validation, writing — review and editing. YJJ: investigation, validation, writing — review and editing. KK: conceptualisation, methodology, software, formal analysis, supervision, funding acquisition, project administration, writing — review and editing. All authors read and approved the final manuscript.

## Acknowledgements

Not applicable.

## Supplementary information

**Additional file 1.** Supplementary methods, tables, and figures: data-integrity audit; seed-level grouped, temporal, and leave-one-pathogen-out results; later-period confusion matrices and cluster-bootstrap intervals; calibration, decision-curve, missingness, and temporal-block analyses; cascade scenarios; model-family robustness of the simulation; Figures S1–S2; and the completed TRIPOD+AI checklist.

## References

- [1] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multi-variable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *J Clin Epidemiol*. 2015;68(2):134–143. doi:10.1016/j.jclinepi.2014.11.010.
- [2] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
- [3] Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51–58. doi:10.7326/M18-1376.
- [4] Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
- [5] Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323. doi:10.2196/jmir.5870.
- [6] Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. *J Med Internet Res*. 2023;25:e48763. doi:10.2196/48763.

- [7] Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138. doi:10.1097/EDE.0b013e3181c30fb2.
- [8] Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54:774–781. doi:10.1016/S0895-4356(01)00341-9.
- [9] Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98(9):683–690. doi:10.1136/heartjnl-2011-301246.
- [10] Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLOS Med*. 2014;11:e1001744. doi:10.1371/journal.pmed.1001744.
- [11] Snell KIE, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol*. 2021;135:79–89. doi:10.1016/j.jclinepi.2021.02.011.
- [12] Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441.
- [13] Takada T, Nijman S, Denaxas S, et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. *J Clin Epidemiol*. 2021;137:83–91. doi:10.1016/j.jclinepi.2021.03.025.
- [14] Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19:453–473. doi:10.1002/1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5.
- [15] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.

Stat Med. 1996;15:361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[16] Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. 2024;384:e074819. doi:10.1136/bmj-2023-074819.

[17] Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data*. 2012;6(4):Article 15. doi:10.1145/2382577.2382579.

[18] Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. *GigaScience*. 2017;6(5):1–9. doi:10.1093/gigascience/gix019.

[19] Bussola N, Marcolini A, Maggio V, Jurman G, Furlanello C. AI slipping on tiles: data leakage in digital pathology. In: *Pattern Recognition. ICPR International Workshops*. Springer; 2021:167–182. doi:10.1007/978-3-030-68763-2\_13.

[20] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91. doi:10.1186/1471-2105-7-91.

[21] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*. 2002;99:6562–6566. doi:10.1073/pnas.102102699.

[22] Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform*. 2014;6:10. doi:10.1186/1758-2946-6-10.

[23] Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328. doi:10.1136/bmj.m1328.

- [24] Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell.* 2021;3:199–217. doi:10.1038/s42256-021-00307-0.
- [25] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns.* 2023;4:100804. doi:10.1016/j.patter.2023.100804.
- [26] Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* 2018;178:1544–1547. doi:10.1001/jamainternmed.2018.3763.
- [27] Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med.* 2019;2:31. doi:10.1038/s41746-019-0105-1.
- [28] Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn.* 2020;2020:151–159. doi:10.1145/3368555.3384468.
- [29] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683.
- [30] Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut learning in deep neural networks. *Nat Mach Intell.* 2020;2:665–673. doi:10.1038/s42256-020-00257-z.
- [31] Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health.* 2020;2:e489–e492. doi:10.1016/S2589-7500(20)30186-2.
- [32] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380:1347–1358. doi:10.1056/NEJMr1814259.

- 874 [33] Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*.  
875 2018;319:1317–1318. doi:10.1001/jama.2017.18391.
- 876 [34] Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. *Neu-*  
877 *roimage*. 2018;180(Pt A):68–77. doi:10.1016/j.neuroimage.2017.06.061.
- 878 [35] Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation.  
879 *J Mach Learn Res*. 2004;5:1089–1105.
- 880 [36] Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation.  
881 *J Am Stat Assoc*. 1983;78(382):316–331. doi:10.1080/01621459.1983.10477973.
- 882 [37] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for  
883 clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI  
884 extension. *Nat Med*. 2020;26:1364–1374. doi:10.1038/s41591-020-1034-x.
- 885 [38] Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interven-  
886 tions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26:1351–  
887 1363. doi:10.1038/s41591-020-1037-7.
- 888 [39] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical  
889 evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat*  
890 *Med*. 2022;28:924–933. doi:10.1038/s41591-022-01772-9.
- 891 [40] Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical ar-  
892 tificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26:1320–1324.  
893 doi:10.1038/s41591-020-1041-y.
- 894 [41] Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imag-  
895 ing (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:e200029.  
896 doi:10.1148/ryai.2020200029.
- 897 [42] Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence  
898 in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6:e240300.  
899 doi:10.1148/ryai.240300.

- [43] Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. *Stat Methods Med Res.* 2018;27(6):1723–1736. doi:10.1177/0962280216668555.
- [44] Little MA, Varoquaux G, Saeb S, et al. Using and understanding cross-validation strategies. Perspectives on Saeb et al. *Gigascience.* 2017;6(5):1–6. doi:10.1093/gigascience/gix020.
- [45] Carlson CJ, Boyce MR, Dunne M, et al. The World Health Organization’s Disease Outbreak News: a retrospective database. *PLOS Glob Public Health.* 2023;3:e0001083. doi:10.1371/journal.pgph.0001083.
- [46] Strack B, DeShazo JP, Gennings C, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *Biomed Res Int.* 2014;2014:781670. doi:10.1155/2014/781670.
- [47] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med.* 2019;38(11):2074–2102. doi:10.1002/sim.8086.
- [48] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837–845. doi:10.2307/2531595.
- [49] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning.* 2006:233–240. doi:10.1145/1143844.1143874.
- [50] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 1947;12:153–157. doi:10.1007/BF02295996.
- [51] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* 2016;74:167–176. doi:10.1016/j.jclinepi.2015.12.005.

- [52] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230. doi:10.1186/s12916-019-1466-7.
- [53] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26:565–574. doi:10.1177/0272989X06295361.
- [54] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2:56–67. doi:10.1038/s42256-019-0138-9.
- [55] Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ.* 2023;380:e071018. doi:10.1136/bmj-2022-071018.
- [56] Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. *BMJ.* 2023;380:e071058. doi:10.1136/bmj-2022-071058.
- [57] Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* 2009;338:b2393. doi:10.1136/bmj.b2393.
- [58] Little RJ, D’Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med.* 2012;367:1355–1360. doi:10.1056/NEJMs1203730.
- [59] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics.* 2012;28(1):112–118. doi:10.1093/bioinformatics/btr597.
- [60] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc.* 2017;24:198–208. doi:10.1093/jamia/ocw042.

- [61] Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform.* 2018;22(5):1589–1604. doi:10.1109/JBHI.2017.2767063.
- [62] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for health-care: review, opportunities and challenges. *Brief Bioinform.* 2018;19:1236–1246. doi:10.1093/bib/bbx044.
- [63] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17:195. doi:10.1186/s12916-019-1426-2.
- [64] Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* 2019;25:1337–1340. doi:10.1038/s41591-019-0548-6.
- [65] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* 2021;3:e745–e750. doi:10.1016/S2589-7500(21)00208-9.

## Supplementary Files

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

- [Additionalfile1.pdf](#)