Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Association between clinical feature encoding and prediction model performance: an equivalence analysis of graded representation collapse in two critical-care databases

Yngve Mikkelsen

yngve.mikkelsen@yahoo.com

Independent Researcher

Research Article

Keywords: research methodology, clinical prediction models, electronic health records, data harmonisation, equivalence margin, measurement heterogeneity, external validation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11362349/v1

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

Additional Declarations: Competing interest reported. Y.M. is Chief Medical Officer of AlgiPharma AS, a clinical-stage biopharmaceutical company. There is no relationship between that role and this work: the study was conceived, conducted and written independently, used only publicly accessible credentialed-access research databases, and AlgiPharma provided no funding, data, input or review. Y.M. declares no other competing interests, financial or non-financial, and has no relationship with this journal in any editorial or advisory capacity.

Abstract

Background

How a clinical feature is encoded is a methodological choice made during data preparation, but its consequences for prediction-model performance are rarely quantified. We estimated how much predictive performance is lost when a categorical electronic health record (EHR) feature encoding is collapsed, against a negligibility margin set before interval estimation.

Methods

Lower-lobe lung-sound fields were analysed in MIMIC-IV v3.1 and compared with the corresponding breath-sound assessment in eICU-CRD v2.0. Within MIMIC-IV, same-timestamp field co-population, categorical support and bilateral agreement were quantified. A feature-level representation audit compared realised supports across corpora, classified cross-corpus relations, and identified irreversible partition changes. On approximately 490,000 at-risk stay-hours per lung-sound field, four encodings were evaluated by five-fold cross-validation with folds grouped by patient for prediction of invasive ventilation within 24 h, using incremental held-out log score and change in area under the receiver operating characteristic curve (AUC).

Results

Of 93,603 MIMIC-IV stays with either lower-lobe field, 93,596 had both; 99.64% of right-lobe entries had a left-lobe entry at the identical chart time, 99.91% of them by the same caregiver_id. Field-specific spelling variants created 9,890 artificial bilateral disagreements; reconciling them raised Cohen's kappa from 0.8704 to 0.8801. The MIMIC and eICU supports contained 15 and 8 strings, including a three-to-one wheeze collapse. The lung-sound field added measurable predictive information (approximately 0.0012–0.0016 nats per observation; ΔAUC 0.019–0.024). In a paired patient-level bootstrap, all 16 log-score contrasts (four outcomes, two model specifications) lay within a 10% margin of the native gain defined before interval estimation; the largest loss bound was 5.2%, and AUC changes were negligible.

Conclusions

Collapsing a categorical EHR feature encoding, including an irreversible three-to-one merge, cost no more than 5.2% of the feature's own predictive contribution across four outcomes and two model specifications. Encoding choices of this kind therefore carry a small, quantifiable cost. The corollary is that performance equivalence cannot by itself establish that two settings represent a predictor equivalently, so predictor provenance requires evidence distinct from task performance.

Background

Electronic health record data are not direct observations of disease. They are produced through clinical and information-system processes that determine what is observed, when it is recorded, and how it is represented. Hripcsak and Albers framed EHR phenotyping as an inverse problem in which patient state must be separated from the process that generated the record [1]. Lasko, Strobl and Stead subsequently formalised site-specific clinical data-generating processes as a central reason probabilistic clinical models may fail to transport between settings [2].

A related clinical prediction literature has examined predictor measurement heterogeneity directly. Pajouheshnia et al. showed that variation in predictor measurement can alter discrimination and transportability [3]. Luijken et al. formalised this problem from a measurement-error perspective [4] and demonstrated in clinical examples that changing measurement procedures can alter calibration, discrimination and overall accuracy to different degrees [5]. Importantly, the broader principle that measurement invariance and predictive invariance are distinct is older than the clinical-AI literature. Millsap proved that the two forms of invariance are not only non-equivalent but inconsistent under realistic conditions, so observed predictive invariance cannot be taken as evidence of measurement invariance [6]. What this literature does not supply is an estimate of how much performance a specific, directly observed encoding change actually costs. To our knowledge, no study has linked a locally generated field-level difference to a real cross-system semantic collapse and then quantified, against a margin set in advance, how far that collapse moves downstream performance.

The size of that cost in routine EHR representation is unresolved. Existing work asks whether representation heterogeneity degrades downstream analysis or whether harmonisation repairs it (see Related work), but it reports the direction of the effect rather than its magnitude against a declared threshold. It does not provide a worked estimate in which a locally generated representation difference is directly observed, is shown to become structurally non-equivalent across systems, and is then shown to cost a relevant downstream performance criterion only a small, bounded amount.

This study provides that estimate. First, paired lung-sound fields in MIMIC-IV are used to localise a representation difference that cannot plausibly originate in patient biology. Second, the same assessment is compared with eICU-CRD, where the released categorical support is coarser and requires a partly lossy semantic mapping. Third, the observed transformation is applied within a fixed prediction task to test whether the lost distinctions are visible to model performance. The objective is not to demonstrate a mechanism of transportability loss. It is to quantify how much predictive performance a documented encoding change costs, and thereby to delimit what performance-based validation can establish about representation.

Related work

Existing EHR representation studies largely treat heterogeneity as a problem whose importance is demonstrated by downstream degradation or by improvement after harmonisation. Rasmy et al. showed that alternative standard terminology representations can produce different predictive results across tasks [7]. Kiser et al. showed that grouping granular EHR features using standard vocabularies reduced the loss of performance during external validation between institutions [8]. More recently, Shi et al. demonstrated that substantial coding differences can remain even after sites adopt a common data model and coding system, and developed statistical methods to detect and harmonise those differences [9]. Data-quality work treats related problems directly: Kahn et al. formalised conformance of recorded values to their allowed formats and value sets within a harmonised EHR data-quality framework [10], and Luu et al. documented loss of information, including discordant LOINC code use and inconsistent units, when laboratory results were exchanged between two institutions [11]. These studies establish that representation matters and that harmonisation can improve portability.

They share a common evidential structure: representational heterogeneity is established as consequential by showing that it changes, or that repairing it improves, a downstream result. What remains unshown is whether a directly observed field-level partition change, linked to a real cross-system semantic collapse, can nevertheless pass a defined downstream prediction criterion with only a small, bounded loss — the case that determines what a passing validation result can be taken to establish. The present study supplies that case.

Methods

Study design and data sources

This was a retrospective methodological study using two de-identified critical-care databases available through PhysioNet credentialed access: MIMIC-IV v3.1 and eICU-CRD v2.0. MIMIC-IV contains data from a single academic medical centre and its ICU module is derived from the MetaVision clinical information system [12, 13]. eICU-CRD is a multicentre critical-care database derived from more than 200 US hospitals, covering admissions from 2014 to 2015, whereas MIMIC-IV v3.1 covers admissions from 2008 to 2022; MIMIC-IV dates are shifted independently per patient for de-identification and eICU-CRD represents event timing as offsets from ICU admission, so original calendar dates cannot be reported [14, 15].

Two empirical questions were separated. The first asked whether the same type of clinical assessment is represented on locally non-identical categorical supports. The second asked whether an observed, partly lossy mapping between supports necessarily produces a detectable change in a specified downstream prediction task. The second analysis was not an external validation of a frozen model: each encoding was fitted and evaluated on the same MIMIC-IV rows so that the information retained by the representation, rather than cross-site case mix or care processes, was the object being compared.

Within-MIMIC representation analysis

The construct was not chosen arbitrarily. Candidate fields were identified by a systematic scan: the 40 chartevents items with the highest event counts, among stays in care units contributing at least 200 stays, were each summarised per stay — as the mean recorded value where at least half of values were numeric, and as a presence indicator otherwise — and all pairwise Pearson correlations between them were examined. The presence indicators for the right and left lower-lobe lung-sound fields emerged as the most strongly correlated pair (r = 0.996), which prompted inspection of their value supports. Field selection was therefore exploratory. All subsequent analyses are conditional on that selection; the study is a worked empirical demonstration rather than an estimate of the prevalence of representation non-invariance. All ICU stays in which either lower-lobe field was documented were included. The extraction covered 432,997,491 charted events and identified 93,603 qualifying stays. Right and left lower-lobe fields were identified by item identifiers 223987 (RLL Lung Sounds) and 223989 (LLL Lung Sounds). Two entries were treated as belonging to the same documentation episode when they had identical charttime within a stay. This definition establishes temporal co-occurrence. The caregiver_id recorded on each chartevents row was compared between the paired entries to test whether the same user documented both fields.

Co-population was quantified as the proportion of right-lobe entries with a same-timestamp left-lobe entry and, separately, as the number of stays containing one field but not the other. The observed categorical supports were compared directly. Bilateral agreement on paired observations was quantified using Cohen's kappa. Where two field-specific strings denoted the same clinical finding, kappa was recomputed after semantic reconciliation of those strings.

Cross-corpus representation comparison

The corresponding nursing breath-sounds assessment in eICU-CRD was identified from the nursing-assessment path. Hospitals contributing at least 50 stays with the assessment were retained, yielding 11 hospitals and 12,048 stays. Right- and left-lower-quadrant entries were paired at identical assessment offsets. Co-population and Cohen's kappa were computed pooled and by hospital. Categorical supports were additionally compared between hospitals within eICU-CRD, using pairwise Jaccard overlap of the realised value sets, a count of values used by only one hospital, and a search for value pairs within Levenshtein distance 2 used by disjoint hospital sets; values occurring fewer than five times at a hospital were dropped before comparison.

The released MIMIC-IV and eICU-CRD categorical supports were compared descriptively. Vendor, era, source implementation and data curation differ between the corpora, so the analysis does not attribute any between-corpus difference to a single organisational or software cause. The resulting supports were then evaluated with the feature-level representation audit described below.

Feature-level representation audit

To separate representation compatibility from task performance, a feature-level audit was applied before downstream modelling. Its inputs were the realised categorical supports for corresponding predictors in each source. The audit comprised four steps: (1) enumerate each support independently; (2) compare values by clinical meaning rather than field name or literal string; (3) classify cross-source relations as exact equivalents, synonyms or relabellings, hierarchical collapses, or no direct equivalent; and (4) determine whether the intended source-to-target mapping is reversible.

One-to-one relabellings were treated as lossless. Many-to-one mappings were treated as partition changes because the original distinction cannot be reconstructed from the coarser support without external information. The audit output was a mapping table identifying preserved and irreversibly lost distinctions. Because the audit uses no outcome data, it establishes representation compatibility rather than task relevance; the latter was tested separately in the downstream experiment.

Downstream representation-collapse experiment

The downstream task used stay-hours in the first 72 h of ICU care with a lung-sound field completed and heart rate, respiratory rate, oxygen saturation, mean arterial pressure and temperature recorded in the same hour. Hours at or after the start of invasive ventilation were excluded. The outcome was initiation of invasive ventilation within 24 h; non-invasive ventilation was excluded. The prediction anchor was the stay-hour rather than the exact assessment timestamp. Because all encoding arms used identical rows and anchors, this limitation does not bias the comparison between encodings, but the baseline model is not given a causal interpretation.

A baseline logistic regression contained heart rate, respiratory rate, oxygen saturation, mean arterial pressure and temperature, standardised within each training fold, with that fold's moments applied to its held-out fold; missing temperature was imputed with the training-fold median. All models were fitted by ridge-penalised logistic regression with a single constant penalty ($\lambda = 10^{-4}$) throughout. The lung-sound field was one-hot encoded and added to the baseline. Five folds were assigned by patient (subject_id), so that no patient contributed observations to both fitting and evaluation. The primary information metric was cross-validated incremental log score: the per-observation held-out log-likelihood gain from adding the lung-sound field. Discrimination was assessed as the change in AUC. Point estimates and bootstrap replicates both used AUC computed on pooled out-of-fold predictions. The experiment was applied to four lung-sound fields: the two lower-lobe fields defined above and the right and left upper-lobe fields (item identifiers 223986 and 223988, d_items labels RUL Lung Sounds and LUL Lung Sounds), all selected by label. Each field was analysed separately on its own at-risk stay-hours, and the same R2 and R3 transformations were applied to each independently; the eICU target support was characterised on the paired lower-quadrant fields. Where a variable was charted more than once within a stay-hour, the value retained was the last by chart time, with ties broken by store time. Ventilation onset was the

earliest procedureevents start time for the item labelled exactly Invasive Ventilation, or Non-invasive Ventilation for that outcome, and in-ICU death was a hospital death time at or before ICU discharge.

Four representations were evaluated on the same rows: R0, the native MIMIC vocabulary; R1, field-specific spelling variants merged without changing the underlying partition (an identity check); R2, the three MIMIC wheeze distinctions collapsed to a single wheeze category; and R3, a mapping to the eight-level eICU support. For R3, exact equivalents were mapped directly, clinically synonymous terms were reconciled, and the three wheeze levels were collapsed. Three rare MIMIC values with no direct eICU equivalent (Bronchial, Tubular and Egophony; together < 0.2% of observations) were assigned by the physician author to the nearest auscultatory category — Bronchial and Tubular to diminished, Egophony to rhonchi — and the full mapping is archived with the code. The whole experiment was repeated with those observations excluded rather than reassigned, as a sensitivity analysis of that choice.

Statistical analysis

All analyses used complete observations for the variables required by the relevant comparison. No imputation was performed for the lung-sound representation or the outcome variables, because the downstream experiment compared alternative encodings on identical rows. Missing temperature was imputed with the median estimated within each training fold and applied to its held-out fold. Cohen's kappa is reported alongside observed and chance agreement where relevant. For the prediction experiment, the difference between each reduced representation and R0 was expressed as absolute change in cross-validated incremental log score, percentage of the R0 gain, and change in AUC. Uncertainty in those contrasts was estimated by a paired patient-level bootstrap with 2,000 replicates. Out-of-fold predictions were generated once per encoding under a single fixed fold assignment shared across all arms; replicates resampled patients with replacement, represented as integer multiplicity weights on their observations, and recomputed both metrics on the stored predictions; for the pooled statistics used here this is exactly equivalent to duplicating the resampled rows. Pairing is essential here: the encodings are applied to identical rows and share nearly all their variance, so an unpaired interval would be dominated by variation common to both arms and would not describe the contrast. Re-splitting within replicates was avoided because it would introduce fold noise into a contrast that is close to deterministic given the folds. A negligibility margin of 10% of the native encoding's own incremental gain was defined before interval estimation, on the log-score scale only. The point estimates of the encoding contrasts were already known from the preceding analysis at that time, so the margin is not a preregistered threshold. No margin was defined for AUC, which is reported descriptively. Intervals are bootstrap percentile intervals.

Two distinct uncertainty analyses were performed, because two different quantities are at issue. Whether the field carries any signal at all was addressed with 20 patient-level outcome permutations; this number was used only to establish the order of magnitude of the null variation, and no formal permutation p-value or tail probability is reported. A synthetic 20-level noise feature was additionally used as an implementation check and produced a small negative held-out gain despite a small positive in-sample gain, as expected for an uninformative predictor.

Robustness across outcomes and model specifications

Two threats to the collapse result were pre-specified and tested. The result could be specific to the chosen target, or it could follow from a baseline too inflexible to represent what the collapsed categories carried. Four outcomes were therefore declared in advance and all are reported: invasive ventilation within 24 h (primary) and within 48 h, non-invasive ventilation within 24 h, and in-ICU death at any point after the observation. In-ICU death was derived from the hospital record of death occurring at or before ICU discharge. Two specifications of the same logistic model were used: the linear baseline of the primary analysis, and a flexible baseline adding quadratic terms and all pairwise products of the five physiologic variables (21 columns against 6), standardised after expansion using training-fold moments. Both are logistic regressions; no second model family was tested. The purpose was not to search for an outcome on which the collapse becomes visible; all declared cells are reported whatever they show.

Tables 2 and 3 were produced by a single implementation. Its fitter was verified against an independent L-BFGS solver at the row count and event rate of the real data, agreeing to within 4×10^{-8} in the coefficients for both specifications at the common penalty, and any fit failing to converge was recorded as a failure rather than reported. Contrasts used the same paired patient-level bootstrap as the primary analysis, with 2,000 replicates and the same predefined margin of 10% of the native gain.

Software and reproducibility

Both source databases are de-identified and distributed under PhysioNet credentialed-access agreements. The analysis involved no patient contact or additional data collection. Analysis was performed in Python using pandas and NumPy. A large language model (Claude, Anthropic) was used to assist with writing and reviewing the analysis code, with statistical cross-checking, and with editing the manuscript text. It is not an author and meets no authorship criterion. All analyses were executed locally on the credentialed data; only aggregate outputs were reviewed with the tool, and no credentialed data were transmitted to it, in keeping with the PhysioNet data-use agreement. All code was reviewed by the author, the logistic fitter was validated against an independent implementation, and the executable scripts are archived. Code availability is stated under Declarations. The experimental logic is summarised in Fig. 1. Reporting was guided by TRIPOD + AI where applicable; the completed checklist is provided as Additional file 1.

Results

A representation difference within one EHR assessment structure

Of 93,603 MIMIC-IV stays in which either lower-lobe field was documented, 93,596 contained both; five contained the right field alone and two the left field alone. At event level, 1,806,527 of 1,813,095 right-lower-lobe entries (99.64%) had a left-lower-lobe entry at the identical chart time. The caregiver_id

recorded on each row was identical for both fields in 1,804,825 of these 1,806,527 pairs (99.91%), with no missing values; the 1,702 discordant pairs comprised 1,531 distinct identifier combinations, only ten of which recurred five times or more. The fields were therefore overwhelmingly co-populated within the same documentation episode and by the same user.

Across paired observations the native union contained 15 distinct strings. Eleven strings were common to both fields. Four strings were field-specific representations of two clinical findings: the right-lower-lobe field contained 'Insp/Exp Wheeze' where the left contained 'Ins/Exp Wheeze', and the right contained 'Pleural friction' where the left contained 'Pleural fricton'. The second pair includes a persistent spelling error. These side-specific patterns are strongly consistent with local field configuration, although the underlying MetaVision configuration object and version history are not available in MIMIC-IV.

The representation difference affected apparent bilateral agreement. There were 9,890 paired observations in which semantically equivalent labels were counted as different strings: 9,794 wheeze pairs and 96 pleural-rub pairs, representing 0.55% of paired observations. Semantic reconciliation reduced the union from 15 native strings to 13 clinical categories and increased Cohen's kappa from 0.8704 to 0.8801. Observed agreement before reconciliation was 0.9263 compared with 0.4315 expected under independence.

The same assessment resolves to a coarser support in eICU-CRD

Among 261,970 right-lower-quadrant eICU entries, 99.76% had a left-lower entry at the identical offset, closely matching the 99.64% co-population in MIMIC-IV. Across the 11 included hospitals, co-population ranged from 0.993 to 0.999. Pooled bilateral kappa was 0.9009, with a hospital range of 0.805 to 0.936. The within-eICU support comparison did not demonstrate hospital-level representation differences. Eight values appeared in total, five were used by all 11 hospitals, and none was confined to a single hospital. The mean pairwise Jaccard overlap of 0.838 is consistent with rarity rather than configuration: the three non-universal values were the three rarest, and the smallest contributing hospital supplied too few observations for them to occur. Because eICU data are integrated into a common warehouse through interface-based transformation and loading, within-corpus uniformity may partly reflect upstream integration or curation. This analysis therefore provides no positive evidence that the between-corpus difference reported below originates at hospital level.

The feature-level representation audit (Table 1) found eight released eICU values compared with 15 native strings in MIMIC-IV. Some differences were lossless semantic relabellings (for example, Crackles to rales and Pleural friction/Pleural fricton to pleural rub). Others were genuinely lossy: MIMIC distinguished inspiratory, expiratory and combined inspiratory/expiratory wheeze, whereas eICU represented these under a single wheezing category. The reverse mapping therefore cannot reconstruct distinctions absent from the eICU support.

Table 1
Clinically interpreted relation between selected MIMIC-IV and eICU-CRD breath-sound values.

MIMIC-IV representation	eICU-CRD representation	Relation
Diminished; Clear; Rhonchi; Absent; Stridor	Corresponding same-named values	Exact equivalent
Crackles	rales	Clinical synonym
Pleural friction / Pleural fricton	pleural rub	Clinical synonym after field-level reconciliation
Insp Wheeze; Exp Wheeze; Insp/Exp Wheeze	wheezing	Hierarchical collapse (3 → 1)
Bronchial; Tubular; Egophony	No direct equivalent	Rare unmatched values (< 0.2% combined)

Representation collapse produced no detectable loss for the primary task

Each lung-sound field was analysed on its own at-risk stay-hours: 490,281 for the right lower lobe, 488,950 for the left lower, 494,535 for the right upper and 494,029 for the left upper, each from about 48,300 patients, with invasive ventilation within 24 h occurring in 3.67% to 3.68%. Across the four fields, the native representation contributed a cross-validated incremental log-score gain of approximately 0.0012 to 0.0016 nats per observation above the physiologic baseline and increased AUC by approximately 0.019 to 0.024. The feature therefore carried task-relevant predictive information.

For the right lower lobe, R0 produced a log-score gain of 0.001169 nats and ΔAUC of 0.0195. The identity representation R1 returned exactly the same values, as required because the design matrix is unchanged up to column labels. Collapsing the three wheeze levels (R2) reduced the gain to 0.001161 nats, an absolute difference of 0.000008 nats (0.7% of the field's gain). Mapping to the eight-level eICU support (R3) reduced it to 0.001159 nats, a difference of 0.000011 nats (0.9%). ΔAUC was 0.0195 in every case.

The other three fields showed the same pattern. All eight point estimates across the four fields were losses, between 0.1% and 1.1% of each field's gain — the direction expected when information is removed. The paired patient-level bootstrap bounds these contrasts rather than merely failing to detect them. Across the four fields, the R2-R0 log-score contrast ranged from − 0.000009 to -0.000002 nats and the R3-R0 contrast from − 0.000018 to -0.000006 nats. Every 95% confidence interval included zero, none extended beyond − 0.000049 or + 0.000016, and all eight contrasts fell entirely within the predefined margin of 10% of the native gain. Contrasts in AUC lay between − 0.00015 and + 0.00008, with intervals contained within − 0.00044 to + 0.00028, against a native contribution of 0.019 to 0.024; the bounded change is thus under 2% of each field's own discrimination. That the field carries signal at all was confirmed separately: across 20 patient-level outcome permutations the incremental log-score gain had mean − 0.000101 and standard deviation 0.000049, and ΔAUC had mean 0.00044 and standard deviation 0.00331, in both cases far from the observed values. For the primary task, no contrast detected

a loss, and every contrast was bounded within the predefined margin. Right-lower-lobe results are summarised in Table 2.

Excluding the three unmatched values, rather than reassigning them, changed no conclusion and produced an additional observation. With Bronchial, Tubular and Egophony removed, R2 and R3 became numerically identical at every field (for the right lower lobe, both 0.000008 nats below R0). This is a structural consequence rather than a coincidence: once those values are absent, collapsing the wheeze levels already yields eight categories that map one-to-one onto the eICU support, so the remaining step is pure relabelling. Among values with direct cross-corpus equivalents, the wheeze collapse is therefore the only partition change; excluding the unmatched values made R2 and R3 identical, so the nearest-category assignments affected no conclusion: with them excluded, every contrast remained within the margin.

Table 2
Downstream performance for alternative encodings of the right lower-lobe lung-sound field.

Representation	Levels	Incremental log score (nats)	Δ vs R0 (Rk − R0)	ΔAUC
R0 native	13	0.001169	—	0.0195
R1 spelling merge (identity)	13	0.001169	0.000000	0.0195
R2 wheeze collapsed	11	0.001161	−0.000008	0.0195
R3 eICU support	8	0.001159	−0.000011	0.0195

Δ is Rk − R0, the same convention as Table 3; *negative values indicate a lower incremental gain for the reduced encoding. Differences were computed before rounding.*

The loss remained bounded across outcomes and model specifications

All sixteen declared contrasts fell within the predefined margin (Table 3). The four outcomes span a sixteen-fold range in event rate, from 0.56% for non-invasive ventilation to 8.9% for in-ICU death, and from 4,658 to 76,032 events. Because the margin was defined after the point estimates were known (see Statistical analysis), sensitivity to its choice was examined. The most extreme lower bound across all sixteen contrasts was 5.2% of the native gain (48-hour ventilation, flexible specification, R3), and fifteen of the sixteen contrasts also satisfy a one-sided non-degradation margin of 5%. No upper bound exceeded 3.0% of the native gain. The bounded-loss conclusion therefore does not depend specifically on the 10% threshold, although one contrast lies marginally outside a 5% criterion.

The flexible specification behaved as intended rather than as a formality: it absorbed part of what the lung-sound field had contributed, reducing the native incremental log score for 24-hour ventilation from 0.001169 to 0.000881 nats (25%) and ΔAUC from 0.0195 to 0.0142 (27%). A specification able to compete with the feature did not make the collapse visible.

Table 3
Collapse contrasts across four outcomes and two model specifications, right lower-lobe field.

Outcome	Model	Events	Contrast	Log-score difference (95% CI)	Margin (± 10%)
Vent 24 h	linear	17989	R2 − R0	−0.000008 (− 0.000028 to 0.000012)	± 0.000117
	linear		R3 − R0	−0.000011 (− 0.000037 to 0.000014)	± 0.000117
	flexible		R2 − R0	−0.000009 (− 0.000028 to 0.000012)	± 0.000088
	flexible		R3 − R0	−0.000008 (− 0.000033 to 0.000015)	± 0.000088
Vent 48 h	linear	24554	R2 − R0	−0.000020 (− 0.000051 to 0.000013)	± 0.000150
	linear		R3 − R0	−0.000027 (− 0.000070 to 0.000012)	± 0.000150
	flexible		R2 − R0	−0.000017 (− 0.000047 to 0.000014)	± 0.000118
	flexible		R3 − R0	−0.000023 (− 0.000061 to 0.000014)	± 0.000118
NIV 24 h	linear	4658	R2 − R0	0.000001 (− 0.000010 to 0.000017)	± 0.000069
	linear		R3 − R0	0.000000 (− 0.000014 to 0.000016)	± 0.000069
	flexible		R2 − R0	0.000002 (− 0.000010 to 0.000018)	± 0.000059
	flexible		R3 − R0	0.000001 (− 0.000013 to 0.000017)	± 0.000059
In-ICU death	linear	76032	R2 − R0	−0.000017 (− 0.000037 to 0.000001)	± 0.000415
	linear		R3 − R0	−0.000045 (− 0.000083 to − 0.000009)	± 0.000415
	flexible		R2 − R0	−0.000016 (− 0.000034 to 0.000003)	± 0.000364
	flexible		R3 − R0	−0.000035 (− 0.000074 to 0.000003)	± 0.000364

Contrasts are Rk − R0, so positive values favour the reduced encoding. The margin column gives ± 10% of the native incremental gain for each outcome and specification; every interval lies inside it.

Twelve of the sixteen point estimates are negative, three are positive and one is zero to six decimal places: the reduced encoding tends to fit marginally worse, as expected when information is removed. One interval excludes zero, in the direction of loss: in-ICU death under the linear specification, R3 − R0 = − 0.000045 (− 0.000083 to − 0.000009), 1.1% of the native gain. With sixteen contrasts and no multiplicity adjustment this interval is descriptive; what matters is its magnitude, which lies well inside the margin. The loss from the collapse is therefore detectable in one of sixteen contrasts and material in none.

Discussion

Principal finding and contribution

This study provides a directly observed EHR example in which representation non-invariance is real, partly irreversible and yet carries only a small, bounded cost for a clinically relevant prediction task. Within MIMIC-IV, persistent field-specific strings demonstrate that part of the observed categorical support is generated by the recording implementation rather than by patient biology. Between MIMIC-IV and eICU-CRD, the same type of breath-sound assessment resolves to non-equivalent supports, including a genuine loss of wheeze granularity. Nevertheless, applying that collapse to a fixed MIMIC-IV prediction task produced no detectable degradation for the primary outcome, and across four declared outcomes the loss never exceeded 5.2% of the field's own predictive contribution.

The general logical conclusion is not novel. Measurement and predictive invariance have long been distinguished in psychometrics [6], and clinical prediction methodology has shown that predictor measurement heterogeneity can have heterogeneous effects across performance metrics [3–5]. The contribution here is empirical and informatic: a naturally occurring representation difference is localised inside a production EHR, linked to a real cross-corpus semantic collapse, and then shown to cost downstream task performance only a small, bounded amount. To our knowledge, prior EHR harmonisation studies have not assembled those three steps in one empirical demonstration.

Against that background, the present result addresses the complementary case: representation may change without producing a warning in the selected performance metric. That matters because validation is task-conditioned. A three-level wheeze distinction that is irrelevant for predicting ventilation within 24 h may still be meaningful for another endpoint, for descriptive inference, or for a later model. Once the target support has collapsed three categories into one, the lost distinction cannot be recovered from the target data, even if current predictive performance is preserved.

Implications for validation evidence

The implication is not that external validation is inadequate or that representational differences are necessarily harmful. External performance evaluation answers an essential question: how well does the model perform in the tested setting and task? The present counterexample shows only that this

evidence cannot simultaneously be treated as proof that predictor representation is invariant. Performance and representation are different evidentiary objects.

Current guidance already points in this direction. PROBAST + AI separates the predictor domain from the analysis domain and asks whether predictor definitions, measurement and preprocessing are applicable to the intended setting [16]. TRIPOD + AI likewise strengthens reporting requirements for predictor definition and measurement [17]. The feature-level audit operationalised here is not a new appraisal framework; it is a narrow pre-model procedure for making those predictor-level questions explicit by comparing realised supports and identifying irreversible partition changes before performance is considered. Provenance and implementation metadata can then be used, where available, to investigate how an identified difference arose.

The practical lesson is therefore narrower than a general call for more harmonisation. Harmonisation should be evaluated in relation to the information required for the intended use. Lossless synonym reconciliation and hierarchical collapse are not the same operation. A mapping can be technically successful and performance-neutral while still being semantically irreversible. Treating performance as the only validation endpoint can obscure that distinction.

Limitations

The field pair was identified by an exploratory scan of frequently charted items rather than by a prespecified sampling frame, so the case is a worked demonstration and not an estimate of how often such differences occur. The downstream experiment concerns one clinical construct. Four prediction targets and two specifications of the logistic model were pre-specified and all are reported (see Results), so the result is not specific to a single outcome or to the tested degree of baseline flexibility. No alternative model family was tested, however, and the finding remains confined to one assessment field. Searching further outcomes until a larger loss appeared would not convert this result into confirmatory evidence; any extension should be specified prospectively.

The underlying MetaVision form configuration is not present in the released MIMIC-IV data. Field-specific supports and persistent field-specific spelling patterns strongly indicate local representation generation, but the physical configuration object and its version history are inferred rather than observed.

Same-timestamp co-population identifies a tightly paired documentation episode. The caregiver_id recorded on each chartevents row was identical for both fields in 99.91% of paired observations, so the two entries were almost always made by the same user; this does not establish a single cognitive act, which the representation argument does not require.

The MIMIC-IV/eICU-CRD comparison is confounded by vendor, era, source implementation and dataset curation. It establishes non-equivalence between the released corpora, not the organisational level at which the difference arose.

Three rare MIMIC values without direct eICU equivalents were assigned by clinical judgement; their combined prevalence was below 0.2%. Excluding them rather than reassigning them changed no conclusion: every contrast remained within the margin. The permutation analysis used 20 patient-level permutations and is descriptive; inference on the contrasts rests on the paired bootstrap, not on it.

Finally, the downstream experiment refits each representation within MIMIC-IV and is therefore not a frozen-model external validation. This is deliberate: the purpose was to test whether a representation difference must announce itself through task-relevant information or discrimination. Actual cross-site deployment adds further sources of shift and is outside the causal scope of this study.

Conclusions

A clinical EHR feature can be represented on structurally non-equivalent categorical supports while preserving predictive performance that is equivalent, on the primary log-score criterion, within a predefined margin. In this study, a naturally occurring field-level representation difference in MIMIC-IV generalised to a real cross-corpus difference in semantic granularity, yet a partly lossy mapping to the eICU support produced no detectable degradation in the primary ventilation prediction and a bounded loss across all declared outcomes. This does not show that representation mismatch causes transportability loss. It shows why successful performance evaluation cannot, by itself, establish representation invariance. Predictor representation, provenance and semantic compatibility therefore require evidence distinct from task performance when those properties matter to deployment or reuse. The feature-level audit used here provides that check before model fitting: enumerate realised categorical supports in each source, compare values by clinical meaning rather than field names, classify mappings as lossless relabellings or partition changes, and record whether the mapping is reversible. Because it requires no outcome data or second model, it can identify an irreversible collapse even when downstream performance changes little or remains within a defined equivalence margin. Model developers, harmonisation teams and appraisers of validation evidence can apply it directly.

Abbreviations

AUC
area under the receiver operating characteristic curve
CI
confidence interval
eICU-CRD
eICU Collaborative Research Database
EHR
electronic health record
ICU
intensive care unit
LLM

large language model
MIMIC-IV
Medical Information Mart for Intensive Care IV
PROBAST
Prediction model Risk Of Bias ASsessment Tool
TRIPOD
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Declarations

Ethics approval and consent to participate

This study used secondary, de-identified data from MIMIC-IV and eICU-CRD obtained through PhysioNet credentialed access under data-use agreements. Collection of the patient information and creation of MIMIC-IV were reviewed by the Institutional Review Board of the Beth Israel Deaconess Medical Center, which granted a waiver of informed consent and approved the data-sharing initiative. eICU-CRD is de-identified to meet the safe harbor provision of the US Health Insurance Portability and Accountability Act, with re-identification risk certified by an independent privacy expert (Privacert, Cambridge, MA; HIPAA certification no. 1031219-2). The present analysis involved no patient contact, intervention, re-identification, or access to directly identifiable information, and required no additional ethics approval or consent.

Clinical trial number

Not applicable. This study is a secondary analysis of existing de-identified critical-care databases and is not a clinical trial.

Consent for publication

Not applicable. The manuscript contains no data from any identifiable individual.

Competing Interests

Y.M. is Chief Medical Officer of AlgiPharma AS, a clinical-stage biopharmaceutical company. There is no relationship between that role and this work: the study was conceived, conducted and written independently, used only publicly accessible credentialed-access research databases, and AlgiPharma provided no funding, data, input or review. Y.M. declares no other competing interests, financial or non-financial, and has no relationship with this journal in any editorial or advisory capacity.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contribution

Y.M. is the sole author. Y.M. conceived and designed the study, wrote the analysis code, performed the analysis, interpreted the results, prepared Figure 1 and Tables 1–3, and wrote the manuscript. Y.M. read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The datasets analysed during the current study are available through PhysioNet to credentialed users under data-use agreements that prohibit redistribution: MIMIC-IV v3.1 (https://doi.org/10.13026/kpb9-mt58) and eICU-CRD v2.0 (https://doi.org/10.13026/C2WM1R). All analysis code supporting the conclusions of this article is available in the ehr-representation-noninvariance repository, https://github.com/yngvemikkelsen/ehr-representation-noninvariance, and archived at https://doi.org/10.5281/zenodo.22883207 (version 1.1.0). The code reproduces every reported result from a local credentialed copy of the two databases. Project name: ehr-representation-noninvariance. Archived version: https://doi.org/10.5281/zenodo.22883207. Operating system: platform independent. Programming language: Python 3.11 or later. Other requirements: uv; dependencies declared inline in each script. Licence: MIT. No restrictions to use by non-academics.

Additional files

Additional file 1. File format: DOCX. Title: TRIPOD + AI reporting checklist. Description: completed TRIPOD + AI checklist with applicability statements for this controlled encoding-comparison study.

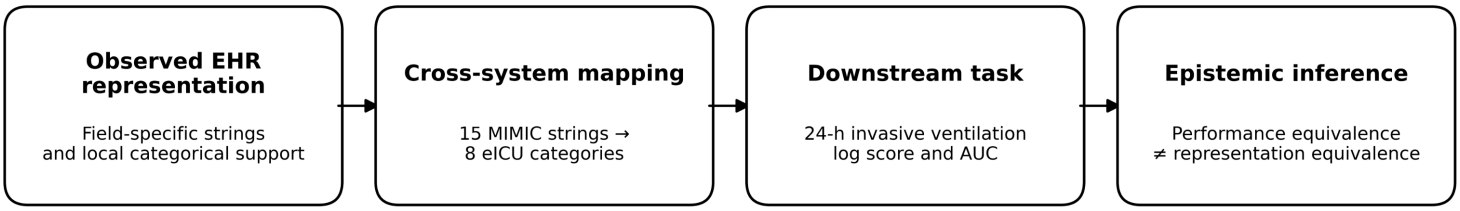
References

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inf Assoc. 2013;20(1):117–21. 10.1136/amiajnl-2012-001145.
2. Lasko TA, Strobl EV, Stead WW. Why do probabilistic clinical models fail to transport between sites. npj Digit Med. 2024;7:53. 10.1038/s41746-024-01037-4.

3. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. J Clin Epidemiol. 2019;105:136–41. 10.1016/j.jclinepi.2018.09.001.
4. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med. 2019;38(18):3444–59. 10.1002/sim.8183.
5. Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH, et al. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. J Clin Epidemiol. 2020;119:7–18. 10.1016/j.jclinepi.2019.11.001.
6. Millsap RE. Measurement invariance, predictive invariance, and the duality paradox. Multivar Behav Res. 1995;30(4):577–605. 10.1207/s15327906mbr3004_6.
7. Rasmy L, Tiryaki F, Zhou Y, Xiang Y, Tao C, Xu H, et al. Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. J Am Med Inf Assoc. 2020;27(10):1593–9. 10.1093/jamia/ocaa180.
8. Kiser AC, Eilbeck K, Ferraro JP, Skarda DE, Samore MH, Bucher B. Standard vocabularies to improve machine learning model transferability with electronic health record data: retrospective cohort study using health care-associated infection. JMIR Med Inf. 2022;10(8):e39057. 10.2196/39057.
9. Shi X, Zhai Y, Yu X, Li X, Hazlehurst BL, Nyongesa DB, et al. Statistical methods to harmonize electronic health record data across healthcare systems: case study and lessons learned. Bioinformatics. 2026;42(3):btag107. 10.1093/bioinformatics/btag107.
10. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. EGEMS (Wash DC). 2016;4(1):1244. 10.13063/2327-9214.1244. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data.
11. Luu HS, Campbell WS, Cholan RA, Edgerton ME, Englund A, Keller A, et al. Analysis of laboratory data transmission between two healthcare institutions using a widely used point-to-point health information exchange platform: a case report. JAMIA Open. 2024;7(2):ooae032. 10.1093/jamiaopen/ooae032.
12. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. 10.1038/s41597-022-01899-x.
13. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet; 2024. 10.13026/kpb9-mt58
14. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. 10.1038/sdata.2018.178.
15. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet; 2019. 10.13026/C2WM1R
16. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, 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. 10.1136/bmj-2024-082505.
17. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.

Figures



The experiment tests whether a real, partly lossy representation change must be visible to task performance.

Figure 1

Experimental logic. The study first identifies an observed EHR representation difference, applies a feature-level representation audit to classify the cross-corpus mapping and its reversibility, then evaluates the resulting representations on a fixed downstream task to delimit what task performance can establish about representation.

Supplementary Files

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

- Additionalfile1TRIPODAIchecklist.docx