BEYOND MODEL LEADERBOARDS: DISENTANGLING DATASET DESIGN AND MODEL PERFORMANCE IN CLINICAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical prediction benchmarks typically fix a dataset and compare models, leaving the consequences of dataset construction underexamined. We introduce a self-contained dataset design system (SDDS) that separates design objects—source, population, features, missingness, and longitudinal structure—from confounding objects governing the model, task, preprocessing, and execution conditions. Individual-object studies and joint experiments over representative factors distinguish object impacts, within-object factor variation, joint factor effects, and changes in model comparisons. Across NHANES, EHRSHOT, and two private health-examination cohorts, covering up to eight model families, features and population generally produce the largest within-source impacts, while missingness has a smaller aggregate impact. Under joint variation, however, features lead sample size in NHANES, whereas sample size leads in EHRSHOT. Small aggregate impacts can also conceal consequential design choices: across the evaluated EHRSHOT missingness configurations, evaluation-only masking produces a task-balanced median AUROC loss of 0.070, compared with 0.002 for all-split masking. In a celiac prediction example, applying 5% missingness within the recent 30-day window reverses the ordering of histogram gradient boosting and a linear model. Dataset design therefore shapes model selection and priorities for data collection; neither follows from sample size or a single benchmark score alone.

1 INTRODUCTION

Clinical prediction research commonly fixes a dataset and compares models (Harutyunyan et al., 2019; Wornow et al., 2023). This supports reproducible comparisons, but leaves a prior question underexamined: how does dataset construction shape the comparative conclusion? Patient selection, feature availability, and observation history shape what models learn and what their scores measure.

This question is particularly consequential in healthcare. Public benchmarks and private clinical cohorts differ in populations, measurements, and recording practices. Choosing or adapting a dataset therefore changes the prediction conditions. Even a small change can alter model selection: applying 5% missingness within the recent 30-day window reverses the ordering of histogram gradient boosting and a linear model in our celiac prediction example (Table 3).

Prior work shows that benchmark selection and data preparation can alter model comparisons (Dehghani et al., 2021; Tschalzev et al., 2024). We study clinically interpretable construction choices through a self-contained dataset design system (SDDS), grounded in Evaluatology (Zhan et al., 2025). SDDS separates design objects from confounding objects governing evaluation conditions. Source constrains four within-source objects: Population, Features, Missingness, and Longitudinal. The analysis proceeds from object impacts to internal factors and joint variation of representative factors, with local cohorts assessing whether priorities recur beyond public benchmarks.

Across four datasets, Features generally has the largest impact, while Population can dominate local multiclass and regression tasks. Missingness has smaller aggregate impacts but contains harmful application-stage choices; Longitudinal sensitivity depends on which history is retained. Joint designs further change factor priorities across datasets. These findings make dataset construction part

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of the evidence for model superiority: a comparative conclusion must specify the data conditions under which it holds.

2 Related Work

Datasets and benchmarks. Dataset construction commonly starts from available sources and intended applications. ImageNet organizes web-retrieved, human-verified images by WordNet categories (Deng et al., 2009). MIMIC integrates and de-identifies routine hospital records (Johnson et al., 2016), while MIMIC-based benchmarks and EHRSHOT define prediction tasks and protocols over existing clinical data (Harutyunyan et al., 2019; Wornow et al., 2023). These efforts emphasize coverage, curation, accessibility, and reproducibility, complemented by guidance on clinical relevance and documentation (Mincu & Roy, 2022; Gebru et al., 2021). They establish resource-specific construction practices, but provide limited guidance for comparing alternative dataset designs. Downstream studies commonly adopt the released data and protocols as references, leaving the dependence of model comparisons on construction choices insufficiently examined.

Dataset analysis and evaluation. Clinical data-quality frameworks characterize completeness, plausibility, and fitness for use (Weiskopf & Weng, 2013; Kahn et al., 2016; Schwabe et al., 2024); dataset-shift studies examine sensitivity to changes in populations and clinical settings (Subbaswamy et al., 2021; Koh et al., 2021; Zech et al., 2018). Feature engineering, label correction, and test-set reconstruction can change scores and model rankings (Tschalzev et al., 2024; Recht et al., 2019; Northcutt et al., 2021). Benchmark-lottery analyses and re-evaluations reveal similar sensitivity to task selection, protocols, and baselines (Dehghani et al., 2021; Musgrave et al., 2020; Dacrema et al., 2019). These studies establish important forms of dataset dependence. SDDS addresses the less developed question of how clinically meaningful construction choices, varied individually and jointly under explicit controls, shape performance and comparative conclusions.

3 Methodology

SDDS treats dataset construction as an experimental component of model evaluation. It specifies the design objects, their feasible dependencies, and the conditions required to interpret performance changes (Zhan et al., 2025).

3.1 Self-contained dataset design system

The system is $\mathrm{SDDS} = (\mathcal{O}, \mathcal{C})$, with five **design objects** (DOs), $\mathcal{O} = \{S, P, F, M, T\}$:

- **Source (S):** source type (public versus private/local).
- **Population (P):** sample size; age, sex, race/ethnicity, and socioeconomic composition.
- **Features (F):** available clinical feature sets, such as laboratory measurements, imaging, diagnoses, procedures, medications, and vital signs.
- **Missingness (M):** mechanism, rate, affected cells or variables, and application to training, validation, or testing data.
- **Longitudinal (T):** observation windows, event budgets, temporal density, minimum events/visits/history length, dropout mode, and retention rate.

Source defines the feasible realizations of the four within-source objects. A feature package is a specified set of clinical variables or event types.

The **confounding objects** (COs) are $\mathcal{C} = \{C_{\mathrm{mod}}, C_{\mathrm{task}}, C_{\mathrm{pre}}, C_{\mathrm{exec}}\}$: model family and training recipe; task, label, and metric definitions; splits, cohort rules, and preprocessing; and execution conditions (budget, seeds, and software/hardware). COs constrain admissible DO settings and are held fixed or stratified when estimating a contrast. Varying the model under common conditions supports model comparisons; varying the source defines separate replication settings.

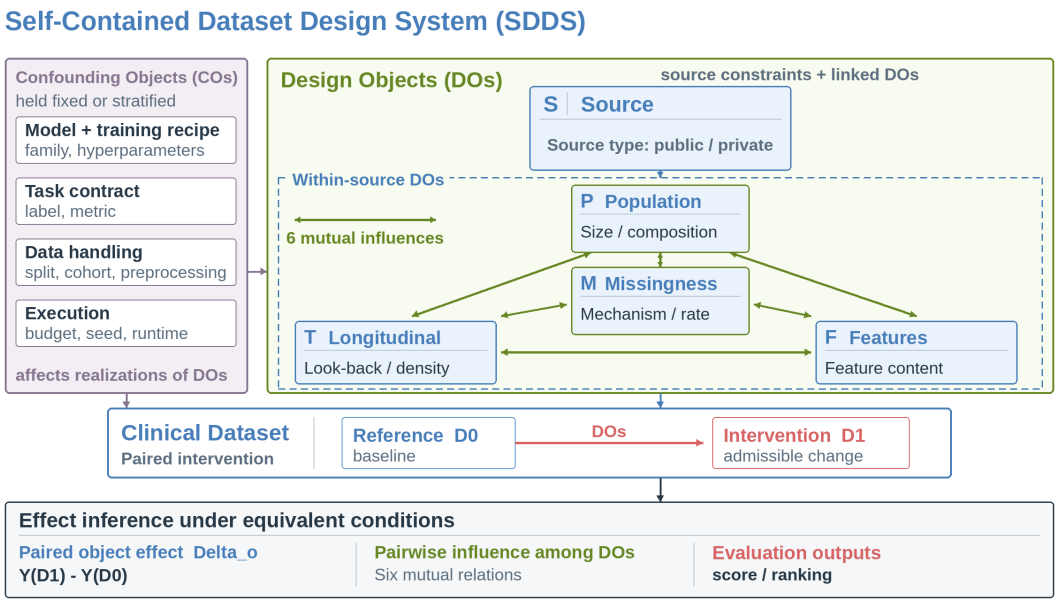

Figure 1: SDDS and its OCM. D_0, D_1 instantiate designs $\mathbf{a}_0, \mathbf{a}_1$. For a fixed model, Δ_o specializes Equation 1 to object o; arrows represent feasible design dependencies.

3.2 Object Causal Model and Interventions

An object causal model (OCM) represents dependencies within SDDS (Figure 1). For example, requiring a measurement may change the eligible population, while shortening history may change both event counts and missingness. An intervention includes the targeted change and any dependent updates required to keep the dataset admissible; these updates are recorded as its *closure*. These controlled changes concern the evaluation system, not clinical treatment effects.

Object-at-a-time (OAT) studies vary attributes within one object's candidate space, retaining the other objects at baseline subject to recorded dependencies. D-optimal selection targets prespecified main-effect terms (Atkinson & Donev, 1992); a baseline and random-audit configurations supplement the selected design. Objects have separate spaces and baselines and are analyzed as parallel studies. **Joint design-of-experiments (DOE) studies** cross representative factors selected from several objects. Interaction estimates concern these representative factors and levels.

3.3 Performance Contrasts and Model Comparisons

Let $\mathbf{a}$ denote an admissible design and $\mathbf{c}_{-m}$ the CO settings other than model and training recipe m. For a fixed source, task, and metric, the performance contrast is

$$\Delta_m(\mathbf{a}_1, \mathbf{a}_0 \mid \mathbf{c}_{-m}) = Y_m(\mathbf{a}_1 \mid \mathbf{c}_{-m}) - Y_m(\mathbf{a}_0 \mid \mathbf{c}_{-m}). \tag{1}$$

For baseline-relative reporting, $\mathbf{a}_0$ is matched within source, experiment, object, task, prediction target, model, and metric. Primary metrics are oriented so that negative contrasts denote degradation. Contrasts are expressed in metric points. Each contrast states whether eligibility changes the test population.

The change in the gap between two models is

$$\Gamma_{m_1,m_2}(\mathbf{a}_1, \mathbf{a}_0 \mid \mathbf{c}_{-m}) = \Delta_{m_1}(\mathbf{a}_1, \mathbf{a}_0 \mid \mathbf{c}_{-m}) - \Delta_{m_2}(\mathbf{a}_1, \mathbf{a}_0 \mid \mathbf{c}_{-m}). \tag{2}$$

A nonzero Γ changes the gap; a strict reversal additionally requires opposite signs of $Y_{m_1} - Y_{m_2}$ in the two configurations. Comparisons use common model support in both configurations.

3.4 Analysis from Objects to Factors

Object impacts. The analysis unit is a realized design point $\times$ task-specific prediction target $\times$ model family within each source, experiment, metric, and split. A design point is a tuple of realized

factor values. Repeated runs are summarized by medians; distinct training/test-size alignments remain separate. Object *impacts* are descriptive summaries: prediction-target-by-model changes are aggregated within umbrella tasks, then across tasks. Analyses are stratified by cohort and task class.

Within-object factor variation. OAT analyses summarize baseline-relative performance changes for the factor levels realized within each object. Because OAT factors are not jointly crossed and some factor combinations are structurally absent, these analyses are treated descriptively rather than as mutually adjusted factor effects.

Joint variation and conditionality. DOE models include representative main effects and supported interactions. Nested ΔR^2 measures the additional variation in evaluation scores explained by an interaction; the reported tests estimate pooled interactions while accounting for task and model family. Separate task- and model-stratified profiles descriptively summarize baseline-relative changes across factor levels. Here ΔR^2 describes the explanatory regression, separately from predictive R^2 on clinical regression tasks. Missingness-stage and temporal-placement comparisons condition on masking rate or retained fraction where supported. Sample–feature comparisons retain the baseline associated with each design; absolute performance trade-offs require a common evaluation population and baseline.

4 Experimental Design

Datasets and tasks. We study two public research datasets, NHANES and EHRSHOT (National Center for Health Statistics, a;b; Wornow et al., 2023), and two private hospital health-examination cohorts, PHE1 and PHE2. The NHANES data come from the 2017–March 2020 pre-pandemic and August 2021–August 2023 releases (National Center for Health Statistics, a;b). Table 1 summarizes the study coverage. Binary tasks include screening, laboratory abnormalities, resource use, and incident disease; multiclass tasks describe severity; regression tasks impute clinical measurements. Panel tasks have multiple targets; task-balanced summaries combine their metric changes. Task rosters and identifiers are given in Appendix A.1.For EHRSHOT, we evaluate 14 structured-EHR prediction tasks and exclude the imaging task, which falls outside the structured-record interventions considered in this study.

Table 1: Cohort coverage. Task counts are binary/multiclass/regression. Model-family counts are roster sizes; eligibility varies by task and configuration.

Cohort	Access	Tasks	Families (OAT/DOE)
NHANES	Public	5/3/7	8/6
EHRSHOT	Public	14/0/0	8/6
PHE1	Private	5/3/7	4/–
PHE2	Private	3/3/7	2/–

Models and evaluation. The public OAT roster contains elastic net, linear models, MLP, Extra Trees, histogram gradient boosting, random forest, FT-Transformer, and TabPFN. The joint DOE uses the remaining six model families, excluding FT-Transformer and TabPFN. Classifier and regressor variants are grouped by model family. Primary metrics are AUROC, macro-AUROC, and R^2 for the three task classes, respectively. Secondary metrics and split identifiers are given in Appendix A. NHANES and EHRSHOT OAT use training, validation, and testing splits; private cohorts use training and testing splits.

OAT configurations. Population, Features, and Missingness are varied in all four cohorts; Longitudinal is varied only in EHRSHOT. Their factors and candidate levels are listed in Table 6. Public designs allocate 40, 24, and 36 D-optimal configurations to the first three objects, each supplemented by one baseline and eight audits. EHRSHOT Longitudinal uses 40 selected configurations, one baseline, and eight audits in a task-independent plan. Evaluated coverage and model exclusions are reported in Appendix B.3, including the implemented TabPFN limit of 50,000 in-context training rows, distinct from the dataset-design sample-size factor.

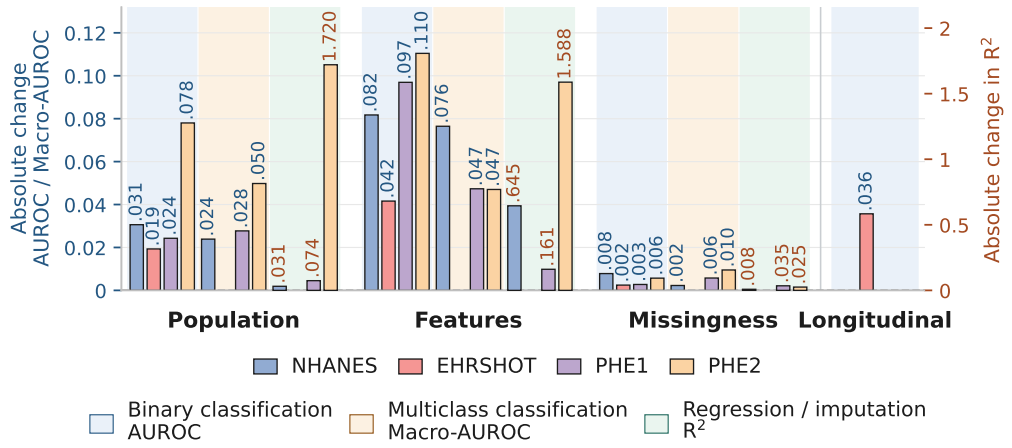

Figure 2: Global object impacts by cohort and task class. Bar colors identify cohorts; pale slots identify task classes, with empty positions unavailable. Bars show absolute task-balanced median changes, so direction is not encoded. AUROC and macro-AUROC use the left axis; R^2 the right.

Joint configurations. NHANES DOE crosses sample size, feature package, and missingness mechanism (133 configurations: one baseline, 120 D-optimal, and 12 random-audit); EHRSHOT additionally varies history window (253 configurations: one baseline, 240 D-optimal, and 12 random-audit). Both use six model families. The induced missingness rate is fixed at 0.2; masking applies to training and validation data in NHANES and to all splits in EHRSHOT. Rate and application-stage effects are examined in the OAT Missingness studies. Factor and interaction analyses focus on the public datasets; local cohorts assess replication at the object level.

5 Results

We examine object impacts, factors within each object, joint effects, and replication across public and local cohorts. Performance changes follow Equation 1.

5.1 Global object-level priorities

Features and Population dominate most settings, but object priorities are not universal. Figure 2 shows that Features has the largest aggregate impact across all task classes in NHANES and PHE1 and in EHRSHOT binary prediction, whereas Population leads multiclass prediction and regression in PHE2. Missingness has the smallest median impact in every available cohort–task comparison. These results establish a recurring first-order pattern—feature coverage and population construction generally matter most—while showing that their relative priority varies by setting.

The exceptions further illustrate this conditionality. Binary Features impacts range from 0.042 to 0.110 AUROC points, while EHRSHOT Longitudinal reaches 0.036, between Features (0.042) and Population (0.019). In PHE2, Population slightly exceeds Features for macro-AUROC (0.050 versus 0.047) and absolute R^2 change (1.720 versus 1.588). Thus, object-level priorities should be interpreted within the evaluated intervention ranges rather than as a universal hierarchy.

5.2 Within-object factor variation

Figure 3 compares the magnitudes of baseline-relative AUROC changes for binary tasks across the factors evaluated in each cohort. The boxes summarize factor-level-by-task medians; they do not show the direction of change or isolate an adjusted factor effect. The corresponding multiclass and regression distributions appear in Figures 7 and 8 in Appendix C.

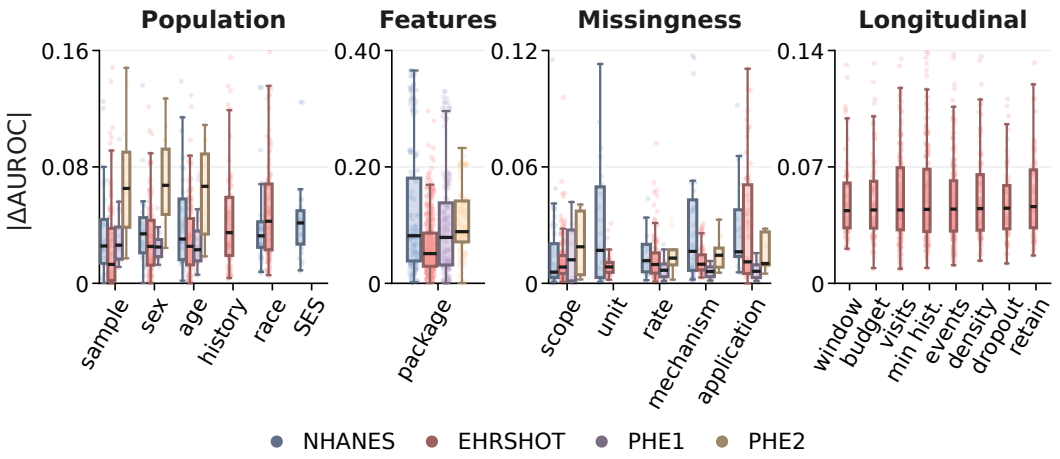

Figure 3: Descriptive absolute baseline-relative AUROC changes for binary tasks by factor and object. Each point is the median $|\Delta\text{AUROC}|$ for a factor level and task across configurations and models; boxes summarize these points. Other factors may vary within a level. Colors identify NHANES, EHRSHOT, PHE1, and PHE2.

Object-level priorities do not reduce to a universal ordering of factors within each object. Figure 3 decomposes the aggregate object impacts in Figure 2 into factor-level changes. Although Features and Population are influential at the object level, their internal factors show substantial overlap and heterogeneity across cohorts.

Population. No single Population factor consistently dominates across cohorts. Sample size, sex, and age show overlapping change distributions where jointly available, while history, race/ethnicity, and socioeconomic factors are evaluated in fewer cohorts. Several factors nevertheless contain observations far above their central ranges. Population importance therefore reflects heterogeneous construction choices rather than a universally dominant factor.

Features. Feature-package changes span the broadest range of factor-level effects across the four cohorts, consistent with the large aggregate features impacts in Figure 2. However, the wide within-factor distributions show that feature importance depends on which information is retained and on the evaluation context, rather than on feature restriction alone.

Missingness. Most missingness interventions produce changes near zero, consistent with its small aggregate impact in Figure 2. However, larger changes occur for particular units, mechanisms, and application stages. Thus, a small aggregate object impact does not imply that all missingness choices are inconsequential; Section 5.4 examines these conditional effects directly.

Longitudinal structure. The eight EHRSHOT factors likewise have broadly overlapping distributions, with no single factor dominating in aggregate. Yet several factors contain substantial changes outside their central ranges, suggesting that temporal sensitivity may depend more on how history is modified than on the identity of a longitudinal factor alone.

5.3 Joint variation of representative factors

There is no common ordering of design factors. Figure 4 shows EHRSHOT level profiles; Table 10 compares both cohorts. NHANES feature availability leads sample size (adjusted $\eta_p^2 = 0.343$ versus 0.237), whereas EHRSHOT sample size leads feature package (0.127 versus 0.086). EHRSHOT history-window length has a much smaller effect (0.0008). Thus, priority differs by source: increasing feature coverage is more consequential in NHANES, whereas increasing sample coverage is more consequential in EHRSHOT.

Average interactions are small for binary AUROC. NHANES's three pairwise interactions add $\Delta R^2 = 0.000622$–0.000880. Five of six EHRSHOT interactions pass $q < 0.01$, but their increments

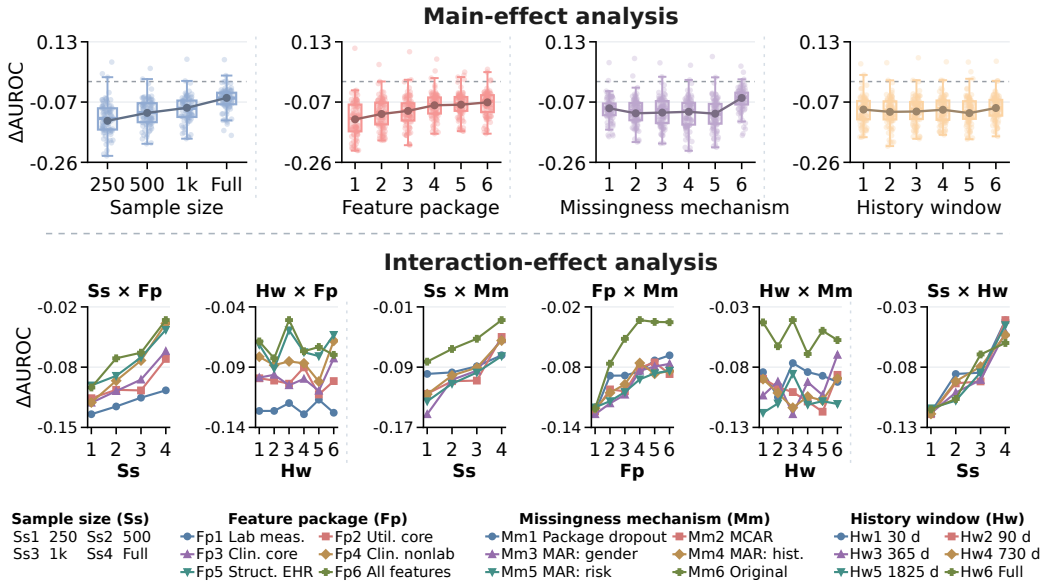

Figure 4: Main- and interaction-effect analysis of the EHRSHOT joint DOE for binary AUROC. Upper panels show level-wise distributions for the four main-effect factors; lower panels show two-factor profiles for the six pairwise interactions. Missingness rate and application are fixed. Corresponding adjusted statistics are reported in Tables 10 and 11.

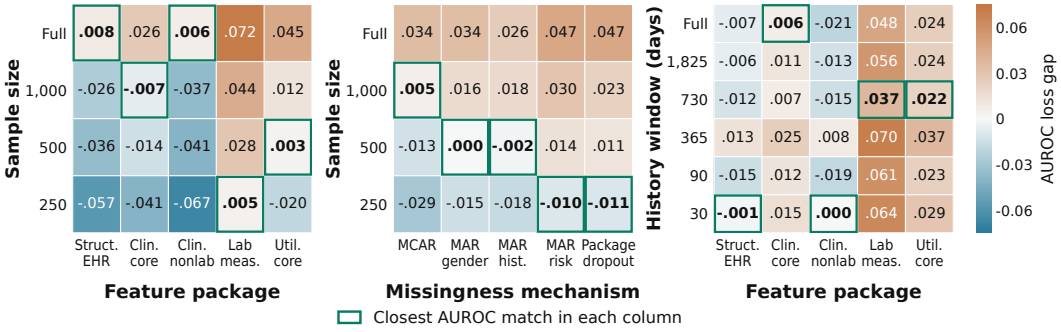

Figure 5: Cross-factor loss analysis of the EHRSHOT joint DOE for binary AUROC. Panels compare feature package with sample size (left), missingness mechanism with sample size (center), and feature package with history window (right). Cells report task-balanced median column-minus-row differences in AUROC loss (points). Green outlines mark the closest observed AUROC-loss match in each column. All available features is the reference feature package and is not shown as an intervention column.

remain 0.001667–0.003738; sample size by feature package is largest (Table 11). Thus, statistical evidence for interaction does not imply a large pooled contribution to explained variation.

Pooled interactions require task-level checks. The small pooled increments do not establish that factor pairs matter equally across clinical tasks. Task-specific joint evaluation remains necessary.

5.4 Implications for Coverage, Missingness, and History

Coverage dimensions do not have a fixed performance-equivalent trade-off. Figure 5 compares AUROC losses across joint-DOE settings. The sample sizes producing the closest observed losses to a given feature restriction vary across feature settings; similarly, the history windows closest to a feature restriction vary with the retained feature package. Therefore, no single sample-size or

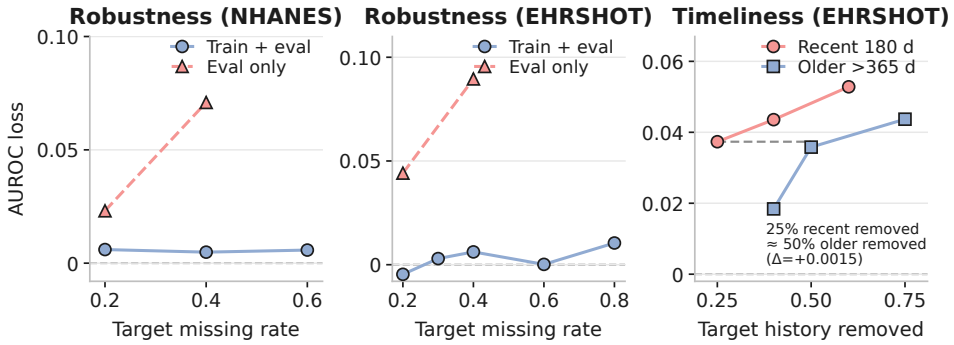

Figure 6: Robustness and timeliness interventions. Left and center: AUROC loss under "Train + eval" versus "Eval only" masking in NHANES and EHRSHOT. Right: removal of recent (180-day) versus older ($>$ 365-day) EHRSHOT history. The horizontal axis is the target removed fraction, equal to one minus the retention rate in Table 6. The marked 25% recent-history and 50% older-history removal settings correspond to retention rates of 75% and 50%, respectively, with an AUROC-loss difference of $\Delta = +0.0015$ points (recent minus older).

history reduction is consistently performance-equivalent to a given loss of feature coverage. These comparisons are descriptive rather than adjusted interaction estimates: in each cell, the column setting is evaluated with the row factor at its reference level, and vice versa, while other DOE factors vary over observed configurations. The closest gaps therefore neither establish equal AUROC nor imply that a fixed sample size is interchangeable with feature or temporal coverage.

Robustness depends on the observation process. Across the evaluated EHRSHOT missingness configurations, "Eval only" (validation-and-testing masking) has a task-balanced median AUROC loss of 0.070, versus 0.002 for "Train + eval" (all-split masking). These are marginal summaries over different rates and mechanisms, not a matched effect of application stage. Figure 6 shows rate-specific losses for these two settings in NHANES and EHRSHOT; in NHANES, "Eval only" masks testing data alone. A separate paired EHRSHOT analysis at rate 0.4 pools testing-only and validation-and-testing masking in one group and training-only, validation-only, and all-split masking in the other. The median paired AUROC contrast between the first and second groups is -0.031 (95% CI $[-0.050, -0.019]$), with 72% of 107 task-by-model units having lower AUROC in the first group. Thus, missingness severity is not characterized by missing rate alone: application stage is associated with substantially different evaluation losses.

Timeliness concerns placement as well as amount. Removing the latest 180 days costs more than removing history older than one year (Figure 6, right). The recent-minus-old gap is -0.021 AUROC (95% CI $[-0.028, -0.015]$) across 107 EHRSHOT task-by-model units, with 82% worse. Notably, removing 25% of recent history produces nearly the same median AUROC loss as removing approximately 50% of older history ($\Delta = +0.0015$). Temporal coverage is therefore not determined by record quantity alone; recent information can be disproportionately predictive.

5.5 Public and local cohorts: recurring priorities and differences

PHE1 retains the Features $>$ Population $>$ Missingness ordering across task classes, while PHE2 Population leads outside binary prediction (Figure 2). Signed contrasts, which Figure 2 does not encode, show a 1.720 R^2 loss for PHE2 Population and a 1.588 gain for Features. PHE1 and PHE2 cover four and two model families, limiting cross-source inference. The local cohorts therefore support recurrence of the broad object-level priorities, while also showing that their ordering is not invariant across sources and task classes.

5.6 Dataset design can reverse model comparisons

For each non-baseline OAT configuration, we compare each model pair's primary metric values with the corresponding values under the matched OAT baseline for the same prediction target within the

same task. We count a reversal when the gap between the models changes sign. Table 2 shows reversals in every available cohort–metric setting, including 28.01% for EHRSHOT AUROC and 29.84% for PHE2 R^2. These rates describe the evaluated task, model, and design spaces, which differ across cohorts. The celiac example in Table 3 shows one such reversal: applying 5% missingness within the recent 30-day window changes the boosting-minus-linear AUROC gap from $+0.057$ to -0.106, giving $\Gamma = -0.163$ under Equation 2. The celiac example illustrates a broader pattern in the evaluated design spaces: comparative model conclusions are conditional on dataset construction.

Table 2: Model-pair ordering reversals relative to the matched OAT baseline. Entries give the percentage (reversals/comparisons) across non-baseline configuration–task–prediction-target–model-pair cells; dashes indicate unavailable task classes.

Cohort	AUROC	Macro-AUROC	R^2
NHANES	**15.30%** (3,312/21,641)	**15.68%** (1,629/10,388)	**18.32%** (7,606/41,524)
EHRSHOT	**28.01%** (16,354/58,380)	–	–
PHE1	**18.28%** (680/3,720)	**16.67%** (372/2,232)	**12.74%** (1,137/8,928)
PHE2	**13.98%** (52/372)	**19.62%** (73/372)	**29.84%** (444/1,488)

Table 3: Pairwise model ordering under an OAT Missingness intervention applying 5% missingness within the recent 30-day window for incident celiac disease prediction (EHRSHOT, C1-15). Bold denotes the higher observed AUROC within each configuration.

Data configuration	Histogram gradient boosting	Linear model
Reference	**0.675**	0.618
5% missingness within recent 30 days	0.617	**0.723**

6 Discussion

SDDS connects dataset-level priorities to the factors that produce them. Features and Population generally have the largest impacts, but their relative importance varies across cohorts and tasks. Missingness and Longitudinal require closer examination of application stage and temporal placement: small aggregate object impacts can coexist with substantial changes under particular configurations. These distinctions make dataset construction part of the evidence for model comparison.

The conclusions are conditional on the evaluated designs. Intervention ranges differ, several OAT factors are not separately estimable, and local cohorts cover fewer model families. Source comparisons reflect multiple construction differences. Although the descriptive profiles show variation across tasks and model families, the reported DOE interaction tests pool these groups and do not establish the magnitude of pairwise factor interactions within individual clinical tasks. The ranking reversal rates and example are descriptive and lack uncertainty estimates; prospective clinical utility and public–local mixing strategies remain untested. Clinical benchmarks should therefore report their design conditions and validate comparative conclusions under the conditions relevant to the intended setting.

7 Conclusion

SDDS evaluates clinical dataset construction through explicit design objects and controlled comparisons. Features and Population generally dominate aggregate impacts, while missingness stage and temporal placement reveal conditional sensitivities. Dataset design can also reverse model orderings. Clinical model comparisons should therefore state the data conditions supporting their conclusions.

Ethics Statement

Data access. NHANES and EHRSHOT were accessed in accordance with their respective data-use requirements. PHE1 and PHE2 consist of de-identified health-examination records made available to the research team through authorized institutional access. Access to the private cohorts was restricted to authorized researchers and conducted in accordance with institutional ethics approval and data-governance requirements.

Reproducibility Statement

Appendix A specifies the cohorts, tasks, splits, and metrics used in the experiments. Appendix B records the intervention factors, configuration coverage, model exclusions, and run provenance, while Appendix C reports additional factor and interaction results. We will release the code and configuration files for the public-data experiments and provide the repository link. The individual-level PHE1 and PHE2 records cannot be publicly released under their institutional data-governance requirements. Their task definitions and aggregate results are documented here, but reproducing private-cohort model training requires authorized access to those records.

AI Use Statement

In this work, we used generative AI tools for language editing, literature search, and code debugging, and to comment on drafts of the experimental design, statistical reporting, and result presentation. We have not used generative AI tools to generate datasets, invent theoretical claims, write proofs, or produce experimental outcomes. All AI-suggested wording, citations, and code edits were manually verified by the authors. Methodological decisions, experiments, statistical analyses, and scientific conclusions were determined by the authors. We take responsibility for the final content, including text, claims, and artifacts produced with generative AI.

References

A C Atkinson and A N Donev. *Optimum Experimental Designs*. Oxford University Press, 08 1992. ISBN 9780198522546. doi: 10.1093/oso/9780198522546.001.0001. URL https://doi.org/10.1093/oso/9780198522546.001.0001.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109, 2019. doi: 10.1145/3298689.3347058. URL https://doi.org/10.1145/3298689.3347058.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery, 2021. URL https://arxiv.org/abs/2107.07002.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1. URL https://doi.org/10.1007/s10994-006-6226-1.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Michael G. Kahn, Tiffany J. Callahan, Juliana Barnard, Alan E. Bauck, Jeff Brown, Bruce N. Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G. Johnson, Siaw-Teng Liaw, Marianne Hamilton-Lopez, Daniella Meeker, Toan C. Ong, Patrick Ryan, Ning Shang, Nicole G. Weiskopf, Chunhua Weng, Meredith N. Zozus, and Lisa Schilling. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs*, 4(1):18, 2016. doi: 10.13063/2327-9214.1244. URL https://doi.org/10.13063/2327-9214.1244.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Diana Mincu and Subhrajit Roy. Developing robust benchmarks for driving forward AI innovation in healthcare. *Nature Machine Intelligence*, 4:916–921, 2022. doi: 10.1038/s42256-022-00559-4. URL https://doi.org/10.1038/s42256-022-00559-4.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision – ECCV 2020*, volume 12370 of *Lecture Notes in Computer Science*, pp. 681–699. Springer, 2020. doi: 10.1007/978-3-030-58595-2_41. URL https://doi.org/10.1007/978-3-030-58595-2_41.

National Center for Health Statistics. NHANES 2017–March 2020 Pre-pandemic. Centers for Disease Control and Prevention data and documentation portal, a. URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020. Accessed: 2026.

National Center for Health Statistics. NHANES August 2021–August 2023. Centers for Disease Control and Prevention data and documentation portal, b. URL https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023. Accessed: 2026.

Curtis Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://openreview.net/forum?id=XccDXrDNLek.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Daniel Schwabe, Katinka Becker, Martin Seyferth, Andreas Klaß, and Tobias Schaeffter. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *npj Digital Medicine*, 7:203, 2024. doi: 10.1038/s41746-024-01196-4. URL https://doi.org/10.1038/s41746-024-01196-4.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2611–2619. PMLR, 2021. URL https://proceedings.mlr.press/v130/subbaswamy21a.html.

Andrej Tschalzev, Sascha Marton, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. A data-centric perspective on evaluating machine learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95896–95930, 2024. doi: 10.52202/

079017-3039. URL https://doi.org/10.52202/079017-3039.

Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013. doi: 10.1136/amiajnl-2011-000681. URL https://doi.org/10.1136/amiajnl-2011-000681.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67125–67137, 2023. doi: 10.52202/075280-2933. URL https://doi.org/10.52202/075280-2933.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683. URL https://doi.org/10.1371/journal.pmed.1002683.

Jianfeng Zhan, Lei Wang, Wanling Gao, Chenxi Wang, Hongxiao Li, Fanda Fan, and Guoxin Kang. *Evaluatology: The Science of Uncovering the Effects*. BenchCouncil Press, Hong Kong, China, 2025. ISBN 978-988-71596-8-1. doi: 10.66834/fr30pm74. URL https://doi.org/10.66834/fr30pm74.

A Cohorts, Tasks, and Evaluation Records

A.1 Task roster and units

Tables 4 and 5 list the tasks. The prefixes C1, C2, and R denote binary classification, multiclass severity, and regression/imputation. In NHANES, PHE1, and PHE2, the glycaemic panel predicts fasting glucose and HbA1c, the renal panel predicts creatinine and uric acid, and the lipid panel predicts HDL-C, LDL-C, total cholesterol, and triglycerides. Task-balanced summaries combine metric changes for these targets within each panel task before aggregating across tasks. PHE1/PHE2 correspond to the legacy archive keys P1/P3. Analyses remain stratified by source.

A.2 Evaluation and provenance

Primary metrics are AUROC (C1), macro-AUROC (C2), and R^2 (R). Secondary metrics are AUPRC for C1; macro-F1, accuracy, and macro-AUPRC for C2; and RMSE and MAE for R. NHANES DOE did not compute macro-F1. Different metric families are never pooled. Model families group variants: logistic and ridge regression are linear; the MLP regressor belongs to MLP. Other models include random forest (Breiman, 2001), Extra Trees (Geurts et al., 2006), FT-Transformer (Gorishniy et al., 2021), and TabPFN (Hollmann et al., 2025).

NHANES and EHRSHOT OAT use training, validation, and testing splits, while local evaluation exports use training and testing splits. Evaluation follows each source's subject grouping, prediction-time boundaries, outcome horizons, and cohort eligibility. NHANES results describe these benchmark cohorts, not survey-weighted populations.

The EHRSHOT note-code history package retains coded events from the note source table and demographic features. It uses event codes and timestamps, not clinical note text.

B Design Factors, Configuration Coverage, and Identification

B.1 Design Factors

Table 4: Public task roster. Family counts denote the experiment roster, not guaranteed coverage of every task–configuration pair.

Task ID	Label	Cohort	Clinical meaning	Models (OAT/DOE)
C1-01	BUA	NHANES	blood/urine abnormality comorbidity	8 / 6
C1-02	CMM	NHANES	cardiometabolic comorbidity	8 / 6
C1-03	PDM	NHANES	diabetes screening	8 / 6
C1-04	PHL	NHANES	hyperlipidemia screening	8 / 6
C1-05	PHT	NHANES	hypertension screening	8 / 6
C1-06	ICU	EHRSHOT	ICU admission	8 / 6
C1-07	LOS	EHRSHOT	length of stay	8 / 6
C1-08	REA	EHRSHOT	readmission	8 / 6
C1-09	ANE	EHRSHOT	anaemia abnormality	8 / 6
C1-10	HKA	EHRSHOT	hyperkalaemia	8 / 6
C1-11	HGL	EHRSHOT	hypoglycaemia	8 / 6
C1-12	HNA	EHRSHOT	hyponatraemia	8 / 6
C1-13	THR	EHRSHOT	thrombocytopenia	8 / 6
C1-14	AMI	EHRSHOT	incident acute MI	8 / 6
C1-15	CEL	EHRSHOT	incident celiac disease	8 / 6
C1-16	HLP	EHRSHOT	incident hyperlipidemia	8 / 6
C1-17	HTN	EHRSHOT	incident hypertension	8 / 6
C1-18	LUP	EHRSHOT	incident SLE	8 / 6
C1-19	PAN	EHRSHOT	incident pancreatic cancer	8 / 6
C2-01	ALT	NHANES	ALT severity	8 / 6
C2-02	BPS	NHANES	blood-pressure severity	8 / 6
C2-03	FGS	NHANES	fasting-glucose severity	8 / 6
R-01	CRE	NHANES	creatinine imputation	8 / 6
R-02	FGL	NHANES	fasting-glucose imputation	8 / 6
R-03	GLY	NHANES	glycaemic panel imputation	8 / 6
R-04	HBA	NHANES	HbA1c imputation	8 / 6
R-05	KID	NHANES	renal panel imputation	8 / 6
R-06	LDL	NHANES	LDL-C imputation	8 / 6
R-07	LIP	NHANES	lipid panel imputation	8 / 6

Table 5: Local task roster. PHE1 has four model families and PHE2 two; neither has joint DOE.

Task ID	Label	Cohort	Clinical meaning	Models (OAT/DOE)
C1-PHE1-01	CMM-PHE1	PHE1	cardiometabolic comorbidity	4 / –
C1-PHE1-02	HEU-PHE1	PHE1	blood/urine abnormality comorbidity	4 / –
C1-PHE1-03	PDM-PHE1	PHE1	diabetes phenotype	4 / –
C1-PHE1-04	PHL-PHE1	PHE1	hyperlipidemia phenotype	4 / –
C1-PHE1-05	PHT-PHE1	PHE1	hypertension phenotype	4 / –
C2-PHE1-01	ALT-PHE1	PHE1	ALT severity	4 / –
C2-PHE1-02	BPS-PHE1	PHE1	blood-pressure severity	4 / –
C2-PHE1-03	FGS-PHE1	PHE1	fasting-glucose severity	4 / –
R-PHE1-01	CRE-PHE1	PHE1	creatinine imputation	4 / –
R-PHE1-02	FGL-PHE1	PHE1	fasting-glucose imputation	4 / –
R-PHE1-03	HBA-PHE1	PHE1	HbA1c imputation	4 / –
R-PHE1-04	LDL-PHE1	PHE1	LDL-C imputation	4 / –
R-PHE1-05	GLY-PHE1	PHE1	glycaemic panel imputation	4 / –
R-PHE1-06	LIP-PHE1	PHE1	lipid panel imputation	4 / –
R-PHE1-07	REN-PHE1	PHE1	renal panel imputation	4 / –
C1-PHE2-01	PDM-PHE2	PHE2	diabetes phenotype	2 / –
C1-PHE2-02	PHL-PHE2	PHE2	hyperlipidemia phenotype	2 / –
C1-PHE2-03	PHT-PHE2	PHE2	hypertension phenotype	2 / –
C2-PHE2-01	ALT-PHE2	PHE2	ALT severity	2 / –
C2-PHE2-02	BPS-PHE2	PHE2	blood-pressure severity	2 / –
C2-PHE2-03	FGS-PHE2	PHE2	fasting-glucose severity	2 / –
R-PHE2-01	CRE-PHE2	PHE2	creatinine imputation	2 / –
R-PHE2-02	FGL-PHE2	PHE2	fasting-glucose imputation	2 / –
R-PHE2-03	HBA-PHE2	PHE2	HbA1c imputation	2 / –
R-PHE2-04	LDL-PHE2	PHE2	LDL-C imputation	2 / –
R-PHE2-05	GLY-PHE2	PHE2	glycaemic panel imputation	2 / –
R-PHE2-06	LIP-PHE2	PHE2	lipid panel imputation	2 / –
R-PHE2-07	REN-PHE2	PHE2	renal panel imputation	2 / –

Table 6: Dataset-design factors and candidate levels in the OAT and DOE experiments. The table lists the union of candidate levels across cohorts and tasks; not every level is applicable or evaluated for every cohort–task combination. Figures summarize the subsets available for the corresponding analyses. N = NHANES, E = EHRSHOT, and P = PHE1/PHE2. OAT varies factors within one design object at a time, whereas DOE jointly varies representative factors selected from different design objects. Semantically equivalent names are harmonized across cohorts. "Original" denotes the unperturbed setting; "all" denotes the complete eligible cohort, available feature package, or recorded history, as applicable.

Exp.	**Object**	**Factor**	**Cohorts**	**Meaning and candidate levels**
OAT	Population	Sample size	N, E, P	Number or fraction of prediction instances, not unique subjects. N: 500, 1000, 2000, 3000, 5000, all. E: 250, 500, 1000, 2500, 5000, 10000, 20000, 50000, all. P: 25%, 50%, 75%, or 100% of the eligible task cohort.
OAT	Population	Age structure	N, E, P	N: original, young/middle/old skewed, old oversampled, ages 40–64 only, ages 65+ only, and balanced age groups. E: the same general levels plus extreme young/middle/old skews. P: observed, balanced, or young/middle/old enriched.
OAT	Population	Sex structure	N, E, P	N: original, balanced, female/male skewed, female only, or male only. E: these levels plus female/male 60% and 90%. P: observed, balanced, female enriched, or male enriched.
OAT	Population	Race/ethnicity structure	N, E	N: original, minority/majority skewing, named-group oversampling, single-group restriction, and balanced major groups. E: original, minority oversampling, and source-coded group oversampling or restriction.
OAT	Population	Socioeconomic structure	N	Original, low-income oversampled, high-income skewed, low-income only, high-income only, or balanced income groups.

Exp.	Object	Factor	Cohorts	Meaning and candidate levels
OAT	Population	History-density structure	E	Original; low-, medium-, or high-history skewed; or low-, medium-, or high-history only.
OAT	Feature	Feature package	N, E, P	All available features; demographics; questionnaires; diet; body, blood-pressure, non-imaging, or combined examinations; imaging; routine, glycaemic, lipid, renal/urine, hematology, liver/biochemistry, or extended laboratory packages; full non-laboratory features; non-laboratory plus named laboratory packages; EHR conditions, medications, procedures, visits/utilization, observations, measurements, notes, devices, deaths, clinical-core combinations, and time-restricted event packages.
OAT	Missingness	Missingness mechanism	N, E, P	MCAR is independent of observed information and values; MAR depends on observed covariates; MNAR-high/low depends on the masked variable itself, preferentially masking high or low values. Original; MCAR; MAR based on age, sex/gender, race/ethnicity, socioeconomic status, risk, or history where available; MNAR based on high or low values; or feature-package dropout. P includes MCAR, MAR-age, and MNAR-high-value.
OAT	Missingness	Missingness rate	N, E, P	N: original, 10%, 20%, 40%, 60%. E: original, 5%, 10%, 20%, 30%, 40%, 60%, 80%, 100%. P: original, 10%, 30%, 50%.
OAT	Missingness	Missingness unit	N, E	N: cell, variable, feature package, subject, or subgroup. E: none at baseline, cell, variable, or feature package.
OAT	Missingness	Missingness scope	N, E, P	N: all, non-demographic, questionnaire, examination, questionnaire plus examination, dietary, high-coverage, low-coverage, or high-missingness features. E: all features, all-history window, condition, drug, procedure, condition–drug–procedure, measurement, non-measurement, observation, note, visit, history density, recent 30 days, or recent 365 days. P: all features or laboratory features.
OAT	Missingness	Missingness application	N, E, P	All splits (training, validation, and testing); training only; validation only; testing only; training and validation; validation and testing.
OAT	Longitudinal	History window	E	Look-back window: 30, 90, 365, 730, or 1825 days, or all recorded history.
OAT	Longitudinal	Event budget	E	Maximum available event count: 25, 50, 100, 200, 500, or 1000 events, or all events.
OAT	Longitudinal	Minimum events	E	0, 1, 2, 5, 20, 25, 50, 75, 100, 150, 200, 250, 500, 750, 1000, 1500, or 2000 events per patient.
OAT	Longitudinal	Minimum visits	E	0, 1, 2, 3, 5, 7, 10, 15, 20, 30, 50, 75, 100, or 150 visits per patient.
OAT	Longitudinal	Minimum history	E	0, 7, 14, 30, 60, 90, 180, 270, 365, 545, 730, 1095, 1460, 1825, or 3650 recorded days.
OAT	Longitudinal	Density distribution	E	Original, low/high-density oversampled, low/middle/high-density only, extreme-high-density only, or balanced low/high density.
OAT	Longitudinal	Dropout mode	E	None, events older than 365 days removed, recent 180-day events removed, random-event dropout, visit dropout, or measurement dropout.
OAT	Longitudinal	Retention rate	E	25%, 40%, 50%, 60%, 75%, 90%, or 100%.
DOE	Population	Sample size	N	500, 1000, 2000, 3000, 5000, or all.
DOE	Population	Sample size	E	250, 500, 1000, 2500, 5000, or all.
DOE	Feature	Feature package	N	All available features, routine laboratory, biochemical laboratory, extended laboratory, full non-laboratory features, or non-laboratory plus biochemistry.

Exp.	Object	Factor	Cohorts	Meaning and candidate levels
DOE	Feature	Feature package	E	All available features, clinical core (Clin. core), clinical non-laboratory (Clin. nonlab), laboratory measurements (Lab meas.), structured EHR (Struct. EHR), or utilization core (Util. core).
DOE	Missingness	Missingness mechanism	N	Original, MCAR, MAR-age, MAR-race/ethnicity, MAR-risk, or feature-package dropout.
DOE	Missingness	Missingness mechanism	E	Original, MCAR, MAR-sex/gender, MAR-history, MAR-risk, or feature-package dropout.
DOE	Longitudinal	History window	E	30, 90, 365, 730, or 1825 days, or all recorded history.
DOE	Fixed conditions	Design controls	N	Original age, sex, race/ethnicity, and socioeconomic structures; 20% missingness; cell-level masking; all-feature scope; and application to training and validation.
DOE	Fixed conditions	Design controls	E	Original age and sex structures; 20% missingness; variable-level masking; all-feature scope; and application to all splits.

In the figures, All features denotes all eligible features for the corresponding cohort and task. Full denotes the complete eligible sample or all recorded history, according to the axis. Package dropout abbreviates feature-package dropout.

PHE1/PHE2 population and missingness dimensions are decoded from their sampled configurations. The observed combinations are not complete factorial products, and the private cohorts are used to validate object-level ordering.

B.2 Design scale and allocation

A design point is a tuple of factor values. Configuration identifiers may be task-scoped: their total is therefore not the number of independent design points. Table 7 reports the exported identifier counts for provenance, while Table 8 reports the planned public OAT allocation. Both joint DOE and OAT have baseline, D-optimal, and random-audit configurations.

Table 7: Recorded component scale. The final column counts configuration identifiers, not independent design points or model-training runs. OAT spaces correspond to within-source objects; DOE has one joint space.

Component	Cohort	Spaces	Tasks	Families	Planned configurations
OAT	NHANES	3	15	8	127 per task
OAT	EHRSHOT	4	14	8	176 per task
OAT	PHE1	3	15	4	127 per task
OAT	PHE2	3	13	2	127 per task
DOE	NHANES	1	15	6	133 per task
DOE	EHRSHOT	1	14	6	253 per task

Table 8: Public OAT allocation. Planned configurations are counted per task, except for the shared EHRSHOT Longitudinal plan. Candidate sets and baselines differ by object.

Cohort	Object	Candidates	Selected	Baseline	Audits	Planned
NHANES	Population	137	40	1	8	49
NHANES	Features	39	24	1	8	33
NHANES	Missingness	170	36	1	8	45
EHRSHOT	Population	97	40	1	8	49
EHRSHOT	Features	84	24	1	8	33
EHRSHOT	Missingness	278	36	1	8	45
EHRSHOT	Longitudinal	520	40	1	8	49

B.3 Coverage and applicability

NHANES evaluates the planned 49/33/45 Population/Features/Missingness configurations per task. EHRSHOT Population evaluates 45–49 of the 49 planned configurations per task; nine tasks have incomplete coverage and 25 task-scoped identifiers were not evaluated by any model. Its Features and Missingness designs evaluate 33 and 45 configurations per task. The Longitudinal plan is task-independent, with 37–49 realized configurations per task. Implemented runs use at most 50,000 training rows as in-context examples ($n_{\text{context}} \leq 50{,}000$), excluding five EHRSHOT laboratory tasks (C1-09 $\sim$ C1-13). This row limit is separate from the dataset-design sample-size factor. Missing or inapplicable cells are not replaced by scores from other models or configurations.

Public and local comparisons use each cohort's available tasks and model families. For PHE2, the binary and multiclass object comparisons each contain six common task-by-model units and are reported descriptively at the object level. Local cohorts are therefore used for descriptive object-level replication, not extrapolation of factor-by-model effects.

B.4 Identifiability of OAT factors

The OAT experiments vary one design object at a time, but the factors within an object are not fully crossed in the realized design. Consequently, a mutually adjusted factor model is not identifiable for all factors. Table 9 reports which factors have sufficient independent support for joint estimation and which remain structurally coupled to other design choices. Terms without an adjusted estimate should therefore not be interpreted as having negligible effects; rather, their independent contributions cannot be separated under the realized OAT configurations. Features contains a single OAT factor, feature package; derived package-grouping and cost-tier labels are not independent factors. We therefore use the OAT analyses descriptively in Section 5.2 and reserve adjusted main-effect and interaction analyses for the joint DOE.

Table 9: Identifiability of binary OAT factors under the realized design. Retained/total gives the number of factors with sufficient support for joint estimation. Terms listed in the final column are not separately estimable because the realized configurations do not provide full-rank support.

Cohort	Object	Identifiable/total	Terms without joint adjustment
NHANES	Population	5/5	None
NHANES	Features	1/1	None
NHANES	Missingness	4/5	Missingness unit
EHRSHOT	Population	4/5	Sex structure
EHRSHOT	Features	1/1	None
EHRSHOT	Missingness	3/5	Missingness rate; Missingness unit
EHRSHOT	Longitudinal	5/8	Retention rate; minimum history; minimum visits

B.5 Run records and contrast interpretation

Each run is linked to its source, task, target measurement where applicable, design point, task-scoped identifier, model family, split, and training/test sizes. Reproducibility records retain factor encodings, selection seeds, preprocessing, execution conditions, and omission reasons. Independent configurations are distinguished from repeated executions of the same configuration. Baseline matching follows Equation 1; all primary contrasts remain in metric points.

Missingness-stage comparisons group training-only, validation-only, and all-split application in one group, and testing-only and validation-and-testing application in the other. Temporal comparisons pair task and model at matched retained fractions or history windows.

C Additional Factor and Interaction Diagnostics

C.1 Binary factor interpretation

Figure 3 shows descriptive absolute AUROC-change distributions; Table 9 documents which OAT factors are separately identifiable under the realized design.

C.2 Multiclass and regression factor variation

Figures 7 and 8 extend the descriptive factor-level comparison in Section 5.2 to multiclass and regression tasks. EHRSHOT has no tasks in these classes.

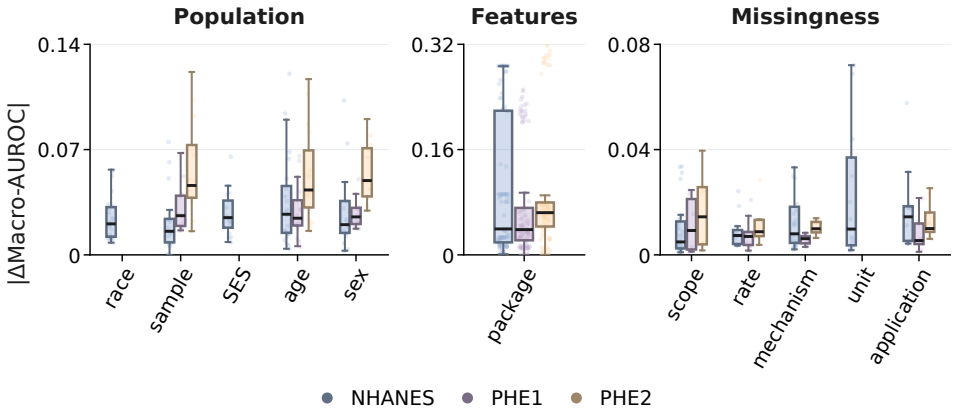

Figure 7: Absolute baseline-relative macro-AUROC changes for multiclass tasks by factor and object. Boxes and points summarize factor-level-by-task change magnitudes; colors identify NHANES, PHE1, and PHE2.

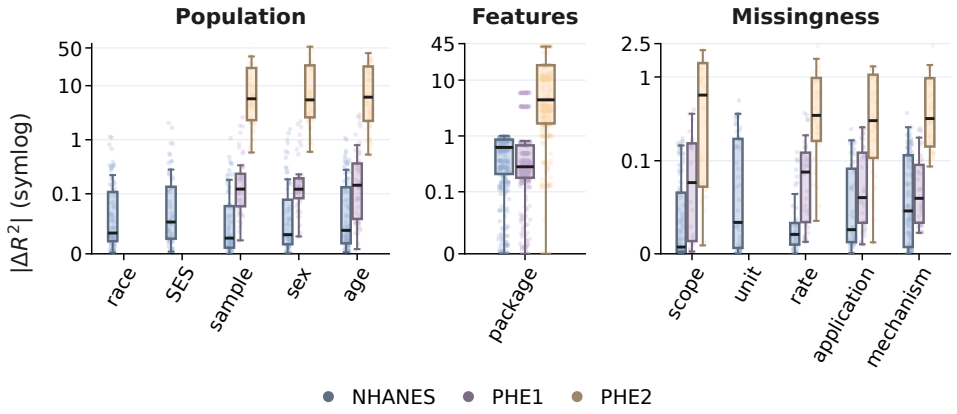

Figure 8: Absolute baseline-relative R^2 changes for regression tasks by factor and object. Boxes and points summarize change magnitudes by factor level and prediction target; colors identify NHANES, PHE1, and PHE2. The vertical axis uses a symmetric logarithmic scale.

C.3 Joint DOE significance tests

Tables 10 and 11 report the ANOVA tests for the representative factors in Section 5.3. Here Ss, Fp, Mm, and Hw denote sample size, feature package, missingness mechanism, and history window, respectively.

C.4 NHANES multiclass and regression joint profiles

Figures 9 and 10 show the joint DOE level profiles for macro-AUROC and predictive R^2. Tables 12 and 13 report their main-effect and pairwise-interaction tests. In both figures and tables, Ss, Fp, and Mm denote sample size, feature package, and missingness mechanism, respectively.

Table 10: Joint DOE main-effect tests for binary AUROC. Partial η_p^2 adjusts for the other DOE factors, task, and model family. The q values use BH correction across seven main-effect tests.

Cohort	Factor	Adjusted η_p^2	F (df$_1$, df$_2$)	p	q	Significant
NHANES	Ss	0.237202	257.010 (4, 3306)	1.65×10^{-192}	2.89×10^{-192}	Yes
NHANES	Fp	0.343278	345.618 (5, 3306)	1.38×10^{-298}	4.84×10^{-298}	Yes
NHANES	Mm	0.061861	43.599 (5, 3306)	1.14×10^{-43}	1.32×10^{-43}	Yes
EHRSHOT	Ss	0.127151	664.560 (3, 13686)	$< 10^{-300}$	$< 10^{-300}$	Yes
EHRSHOT	Fp	0.085711	256.601 (5, 13686)	5.31×10^{-263}	1.24×10^{-262}	Yes
EHRSHOT	Mm	0.050711	146.221 (5, 13686)	1.06×10^{-151}	1.49×10^{-151}	Yes
EHRSHOT	Hw	0.000842	2.305 (5, 13686)	0.04193	0.04193	Yes

Table 11: Binary AUROC joint DOE interaction tests. Each pair extends the additive model for main effects, task, and model family. BH correction covers all nine pairs.

Cohort	Factor pair	ΔR^2	F (df$_1$, df$_2$)	p	q	Significant
NHANES	Ss×Fp	0.000780	0.303 (20, 3286)	0.9988	1.000	No
NHANES	Ss×Mm	0.000880	0.342 (20, 3286)	0.9972	1.000	No
NHANES	Fp×Mm	0.000622	0.193 (25, 3281)	≈ 1.000	1.000	No
EHRSHOT	Ss×Fp	0.003738	5.235 (15, 13671)	1.40×10^{-10}	1.26×10^{-9}	Yes
EHRSHOT	Ss×Mm	0.001667	2.327 (15, 13671)	0.00256	0.00575	Yes
EHRSHOT	Ss×Hw	0.001254	1.749 (15, 13671)	0.03578	0.05368	No
EHRSHOT	Fp×Mm	0.002265	1.897 (25, 13661)	0.00441	0.00794	Yes
EHRSHOT	Fp×Hw	0.002655	2.226 (25, 13661)	0.000413	0.00139	Yes
EHRSHOT	Mm×Hw	0.002638	2.211 (25, 13661)	0.000462	0.00139	Yes

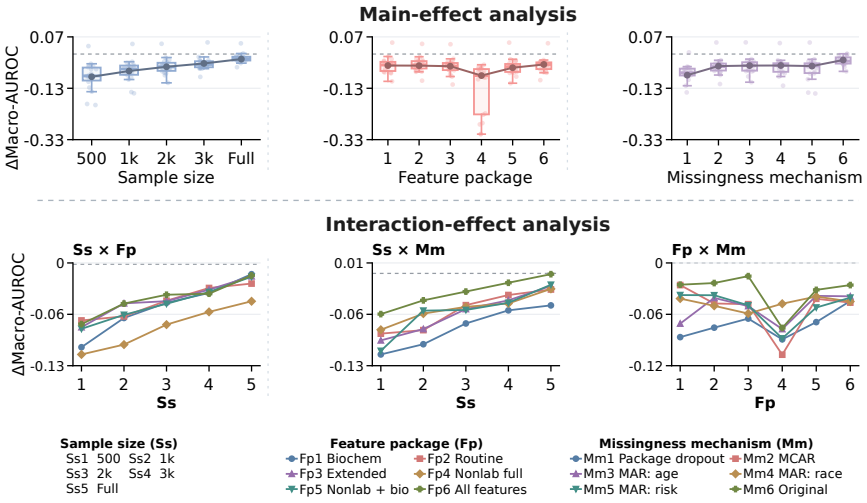

Figure 9: NHANES joint DOE level profiles for multiclass macro-AUROC. Upper panels show baseline-relative changes across sample size (Ss), feature package (Fp), and missingness mechanism (Mm); lower panels show the three pairwise level profiles.

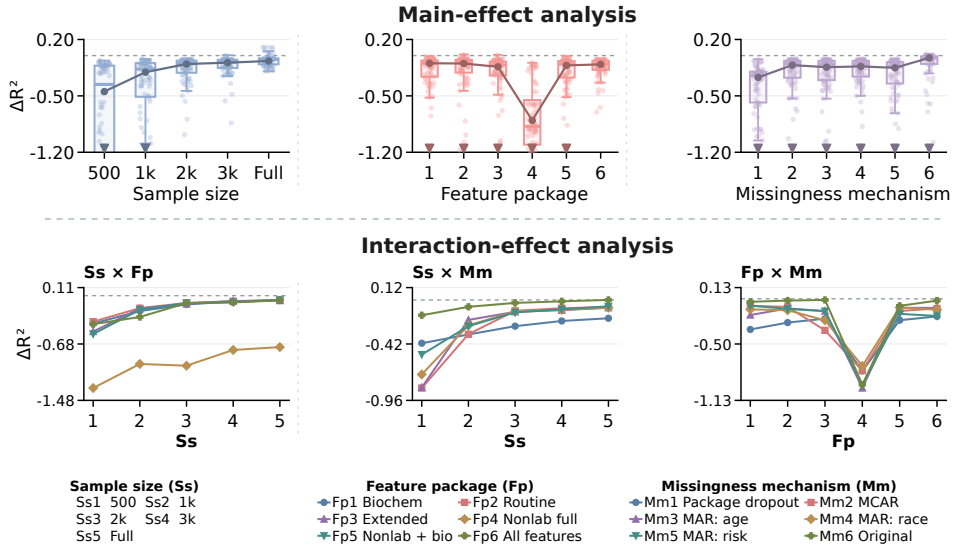

Figure 10: NHANES joint DOE level profiles for regression R^2. Upper panels show baseline-relative changes across sample size (Ss), feature package (Fp), and missingness mechanism (Mm); lower panels show the three pairwise level profiles. This predictive R^2 differs from the explanatory ΔR^2 used to assess interactions.

Table 12: NHANES joint DOE main-effect tests for the other primary metrics. Partial η_p^2 adjusts for the other two factors, prediction target, and model family. BH q values are calculated separately across the three factors of each metric.

Metric	Factor	Partial η_p^2	F (df_1, df_2)	p	q	$q < 0.05$
Macro-AUROC	Ss	0.207262	129.156 (4, 1976)	4.51×10^{-98}	6.76×10^{-98}	Yes
Macro-AUROC	Fp	0.316364	182.885 (5, 1976)	2.68×10^{-160}	8.04×10^{-160}	Yes
Macro-AUROC	Mm	0.046295	19.184 (5, 1976)	1.10×10^{-18}	1.10×10^{-18}	Yes
R^2	Ss	0.008806	17.6825 (4, 7961)	1.84×10^{-14}	5.53×10^{-14}	Yes
R^2	Fp	0.001256	2.00211 (5, 7961)	0.0751	0.1126	No
R^2	Mm	0.001083	1.72623 (5, 7961)	0.1248	0.1248	No

Table 13: NHANES joint DOE pairwise-interaction tests for the other primary metrics. Each pair is added separately to the additive model containing all three factors, prediction target, and model family. ΔR^2 is the increment in explanatory fit; BH q values are calculated within each metric's three pairs.

Metric	Factor pair	ΔR^2	F (df_1, df_2)	p	q	$q < 0.05$
Macro-AUROC	Ss×Fp	0.000542	0.111265 (20, 1956)	1.0000	1.0000	No
Macro-AUROC	Ss×Mm	0.000980	0.201272 (20, 1956)	0.99995	1.0000	No
Macro-AUROC	Fp×Mm	0.001467	0.240766 (25, 1951)	0.999963	1.0000	No
R^2	Ss×Fp	0.002251	0.923977 (20, 7941)	0.555887	0.555887	No
R^2	Ss×Mm	0.004292	1.76495 (20, 7941)	0.018759	0.056277	No
R^2	Fp×Mm	0.003729	1.22554 (25, 7936)	0.201778	0.302667	No

C.5 TASK- AND MODEL-STRATIFIED AUROC PROFILES

We summarize 13,723 EHRSHOT joint-DOE design rows covering 14 binary tasks and six model families, restricted to the four common sample-size levels: 250, 500, 1,000, and the task-specific full cohort. For each task–model-family combination and each focal factor level, we first take the median baseline-relative AUROC change over the sampled background configurations. Task profiles then take the median across model families; model-family profiles take the median across tasks, giving each task equal weight. All plotted levels share the same task and model-family support, although the sampled background configurations may differ.

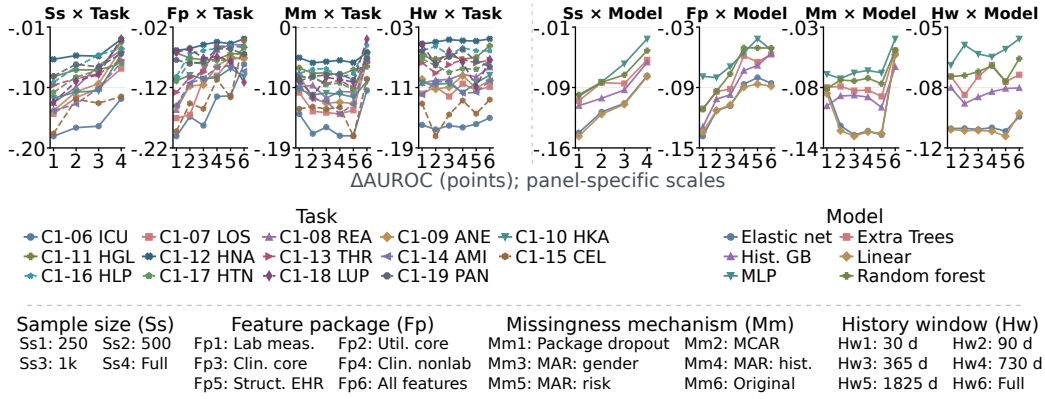

Figure 11: Task- and model-stratified descriptive AUROC profiles under joint variation in EHRSHOT. The left four panels show task profiles and the right four show model-family profiles. Ss, Fp, Mm, and Hw denote sample size, the feature configuration, missingness mechanism, and history window. Fp retains the manuscript's feature-package display label; the underlying EHRSHOT DOE factor is recorded as `module_family`. The horizontal axes index the levels listed below the panels; vertical axes show baseline-relative AUROC changes in metric points, with panel-specific ranges. Full denotes the task-specific full cohort or all available history, respectively. Lines connect level summaries for readability and do not imply a continuous scale for categorical levels. These descriptive summaries are not adjusted interaction estimates or significance tests.

Sample-size endpoint improvements are directionally consistent across task and model-family profiles. Extending the history window from 30 days to all available history increases the summarized AUROC change in nine tasks and decreases it in five. Feature-configuration responses also vary across tasks. These patterns do not quantify task-specific pairwise factor interactions or establish statistically significant differences between profiles.