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.

Closing the Loop on Automated Feature Engineering: A Heart-Failure Pilot

Taimoor Awan
Nimblemind AI
Michelle Hu
Nimblemind AI
Sajeev Singh
Department of Health Informatics, Rutgers University Newark
Piyum Zonooz
Nimblemind AI
Navin Kumar
navin@nimblemind.ai

Nimblemind AI
Priyadarshini Kachroo
Department of Health Informatics, Rutgers University Newark

Research Article

Keywords: electronic health records; heart failure; risk prediction; feature engineering; clinical artificial intelligence; multi-agent systems; evidence-linked extraction; recursive self-improvement

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11816550/v1

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

Additional Declarations: The authors declare no competing interests.

Closing the Loop on Automated Feature Engineering: A Heart-Failure Pilot

Taimoor Awan*, Michelle Hu*, Sajeev Singh†, Piyum Zonooz*, Navin Kumar*, Priyadarshini Kachroo†
*Nimblemind
†Department of Health Informatics, Rutgers University Newark

***Abstract*—Heart-failure risk models draw on variables scattered across diagnoses, medications, laboratory results and procedures. The features built from these records are made by hand for each dataset, with no record of the evidence behind them. We evaluate the Nimblemind Multi-Agent System (nMAS), whose recursive self-improvement (RSI) loop proposes features and model weights from published knowledge. Each predictor mined from the heart-failure literature is built as a feature and kept only if it improves a fixed model on the hospital's own patients. The loop was applied to a synthetic cohort of 10,000 records, 31,427 MIMIC-IV patients and 26,168 patients of a United States health system, predicting death within 365 days of a first heart-failure diagnosis. Across four experiments, five iterations of the loop raised holdout AUROC by +0.008 to +0.047 over cleaned baselines. The accepted features were markers that cardiologists generally use, among them estimated glomerular filtration rate, natriuretic peptides, and iron and liver tests. Each accepted feature carries the published sentence, direction and cutoff it was built from, so a clinician can check the feature before it enters a model used for risk stratification, targeting of guideline-directed therapy or timing of referral for advanced therapies. Run inside the health system with no record leaving the institution, the loop raised holdout AUROC from 0.708 to 0.755, and better ranking of patients by risk supports each of these uses.**

***Index Terms*—electronic health records; heart failure; risk prediction; feature engineering; clinical artificial intelligence; multi-agent systems; evidence-linked extraction**

I. Introduction

Before a clinical prediction model can be trained, routinely collected records have to be turned into analysis-ready variables, and data scientists report that 39% to 45% of their time goes to that step [1]. The step is hardest where predictors are scattered, and heart failure is such a case. It affects an estimated 6.7 million adults in the United States [2]; its prognostic variables sit in diagnoses, medications, laboratory results, procedures and imaging; hundreds of risk models reuse a small core of hand-chosen predictors, and fewer than a quarter are externally validated [3]. Existing language-model feature engineering proposes features from column descriptions or model knowledge [4], cites no evidence, and is evaluated against raw tables rather than against a cleaned clinical baseline. This study asks whether features proposed from the published literature, built automatically, and kept only when they improve prediction on the hospital's own patients, improve one-year mortality prediction in heart failure.

The Nimblemind Multi-Agent System (nMAS) [5], [6] takes a clinician's request, checks that it can be answered from the available tables, and routes it to a pathway such as feature extraction for narrative documents or feature engineering for structured tables. The feature-engineering pathway includes a recursive self-improvement (RSI) loop (Fig. 1): an agent proposes candidate predictors from the literature, builds them, tests them against the current model, and starts the next proposal from the improved table, in the manner of recent self-improving systems [7], [8]. The loop is the component this study evaluates, for two reasons. Each feature it keeps carries the sentence, direction and cutoff it was built from, so the model can be checked by a clinician rather than trusted on a column name. The gate runs on the hospital's own data, so a published finding enters the model only where it holds locally.

Related work. Language-model feature engineering such as CAAFE [4] generates features from column names and the model's own knowledge; feature-selection methods such as Boruta [9] rank columns that already exist; published heart-failure scores such as the Seattle Heart Failure Model and MAGGIC [10], [11] fix a predictor set in advance. None ties a feature to a published sentence or tests it on the local cohort before admitting it; the RSI loop does both.

The RSI loop is used because the right features cannot be chosen in advance: which of the 299 candidate concepts helps depends on the cohort and on what the baseline already holds, so each candidate is built and tested on the hospital's own data and kept only if it passes. Whether such a loop improves prediction on real hospital records beyond a cleaned clinical baseline has not been tested; this pilot on three cohorts is that test, and the rest of nMAS depends on its answer.

In this pilot study of four experiments on three heart-failure cohorts, the RSI loop proposed features, model weights from published knowledge and revisions to its own rules, and the same test was applied before anything was kept. The loop was measured against the same candidate features added without the loop, features chosen by a general-purpose language model, and a stricter baseline.

II. Methods

RSI loop. Recursive self-improvement means that the system proposes a change to its own feature table, tests the change against the current model, keeps it only if the test passes, and starts the next proposal from the improved table. As shown in Fig. 1, nMAS routes each request to a pathway such as feature extraction or feature engineering; the RSI loop of the

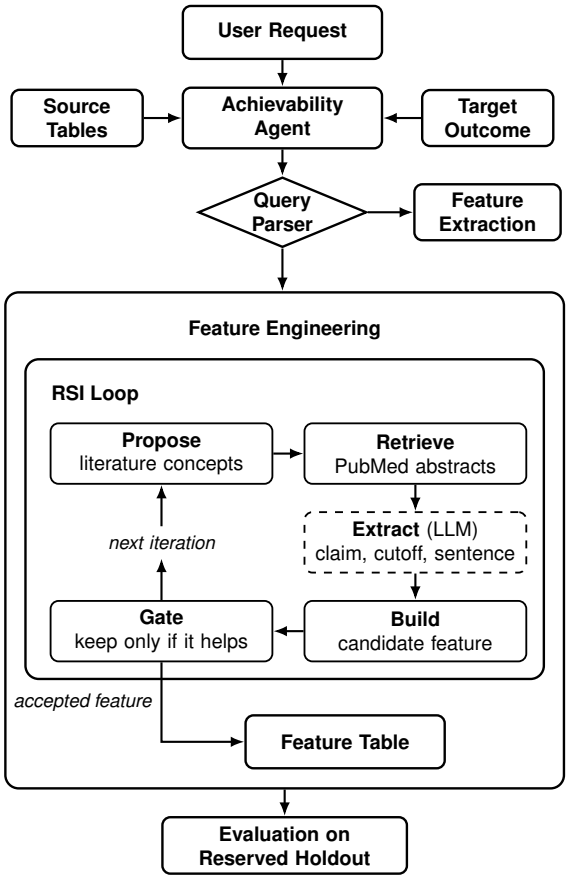


Fig. 1. The Nimblemind Multi-Agent System (nMAS). This study evaluates the RSI loop inside the feature-engineering pathway; the dashed box is the only language-model call, which reads published abstracts and never receives patient data.

feature-engineering pathway is evaluated here. Starting from a cleaned baseline table, each iteration runs five steps:

1) *Propose:* sample ten of 299 MeSH concepts, harvested from heart-failure prognosis abstracts in 26 cardiology journals, that match the available data.
2) *Retrieve:* fetch up to eight eligible PubMed abstracts per concept.
3) *Extract:* a large language model (LLM) reads each abstract and returns the measurement, cutoff, direction and supporting sentence, without seeing column names.
4) *Build:* each retained claim becomes a candidate feature.
5) *Gate:* on five fixed development folds, a candidate is accepted only if it improves the paired log-loss of a logistic model, beats permuted copies of itself, and is stable under bootstrap resampling.

Accepted features stay in the table, and the next iteration starts from it. Patients were split once into a development set (80%) and a reserved holdout (20%). Each loop's accepted set was scored once on the holdout as the change in area under the receiver operating characteristic curve (AUROC) over the baseline with a 95% bootstrap confidence interval (CI); the holdout was never used for sampling, extraction or acceptance. A second analysis measured how much of the full laboratory panel's gain the literature-derived features recover from a table of demographics and diagnoses only.

Experiments. Hospitals run prediction models in different environments: some can send de-identified data to a frontier model, many must keep every step on their own infrastructure [12], and a public benchmark is needed so that results can be reproduced. Four experiments were run: Experiment 1 used a synthetic cohort to develop the loop with a frontier-scale open-weight model, DeepSeek-V4-Flash [13], and to build the corpus of retrieved abstracts that the later experiments reuse. Experiment 2 used MIMIC-IV, an openly available benchmark, so that credentialed researchers can reproduce the results. Experiment 3 used a health-system cohort analyzed inside the institution. Experiment 4 returned to MIMIC-IV to test the loop against a stricter baseline and to apply weight-space programming, in which a language model reads and writes the weights of a model directly, with no optimizer [14], [15]; here the loop proposed the weights of the prediction model from published knowledge. The outcome, death within 365 days of the first heart-failure diagnosis, is the horizon of established heart-failure risk models [10], [11] and can be derived from structured vital-status fields in every source used here. Each loop ran with budgets of 1, 3 and 5 iterations to test whether gains accumulate; five was the largest budget likely feasible on a hospital's compute allocation [16].

Experiment 1: synthetic cohort. The synthetic cohort holds 10,000 heart-failure patient records generated similar to the 500-record cohort of [6], from de-identified cases of a single United States hospital system. The records are spread across nine EHR tables (demographics, diagnoses, medications, procedures, laboratory results, imaging and social history). The outcome is death within 365 days of the first heart-failure diagnosis. The baseline is a cleaned 132-column table covering demographics, diagnoses, medications, procedures, ejection fraction and seven laboratory tests. Extraction used live PubMed retrieval, and the loop ran at 1, 3 and 5 iterations with five seeds each (15 runs).

Experiment 2: MIMIC-IV. Experiment 2 evaluates the same task on MIMIC-IV v3.1 [17], [18], [19], a de-identified record of patients admitted through the emergency department or an intensive care unit at Beth Israel Deaconess Medical Center, Boston, between 2008 and 2022. Adults (age $\geq$ 18 years) were included at their first hospital admission carrying a heart-failure diagnosis (ICD-10 I50, I11.0, I13.0 or I13.2; ICD-9 428, 402.x1 or 404.x1/x3). Patients were split 80/20 into development and holdout sets. The outcome is death within 365 days of the index admission date, including deaths during that admission, taken from hospital records linked to state death records; because MIMIC-IV follows every patient for one year after their last discharge, the outcome is observed for all included patients. Laboratory, medication and outpatient records were limited to those dated up to the end of the admission day. The baseline holds age, sex, race, ethnicity and preferred language, and the most recent outpatient body-mass index, blood pressure and weight. It adds 15 comorbidity flags

and a diagnosis count from ICD codes of prior admissions and the index admission, excluding codes that record the outcome or end-of-life care, and 16 medication-class indicators. It also holds the most recent NT-proBNP, creatinine, potassium, sodium, urea nitrogen, hemoglobin and glucose values. Candidate features draw on all laboratory results, outpatient measurements (including estimated glomerular filtration rate, eGFR), prescriptions grouped into pharmacologic classes, and procedure codes. All loops read the abstract set retrieved in Experiment 1, and each loop extracted its own claims with Qwen2.5-7B-Instruct [20]. The loop ran at 1, 3 and 5 iterations with ten seeds each (30 loops). Gains were also measured with a boosting model.

Experiment 3: health-system cohort. The third cohort is an electronic health record export from RWJBarnabas Health, a New Jersey health system, with a data cutoff in February 2026. Extraction ran inside the institution: Qwen2.5-7B-Instruct read the Experiment 1 abstracts on the site's own hardware, and no record left the institution. Patient flow is given in Appendix Table A1; the cohort was not restricted by age. The baseline holds 23 columns: demographics, vital signs, ten comorbidity flags and an indicator of presence in the diagnosis extract. Laboratory values, medications and procedures were excluded from the baseline because the available derived versions use records dated after diagnosis. The loop ran at 1, 3 and 5 iterations with ten seeds each (30 loops). Gains were measured with a boosting model.

Experiment 4: RSI loop with written weights. Experiment 4 first reran the loop on MIMIC-IV under stricter conditions. Baseline columns with missing values received an indicator that a value was recorded, as the loop's features already had; this was the corrected baseline. Patients were split into development and holdout sets with a new seed, and every result of Experiment 4 used this second holdout. The loop of Experiment 2 was rerun with ten seeds, a corrected loop that scored every eligible concept in each iteration was scored once, and a lasso logistic regression that selected laboratory tests served as a comparator. Experiment 4 then applied weight-space programming within the loop's cycle of proposing, testing and keeping. Weight-space programming was introduced for small neural networks, whose weights a language model wrote directly from a natural-language description of the task, with no optimizer [14], [15]; the coefficients of a logistic model are a simpler case. A written weight is a coefficient set from published knowledge rather than fitted to patients. Written weights came from the ratios reported in the extracted claims and from a frontier model, Claude Fable 5.2, which wrote every weight of the corrected loop's final model from a description of its columns, with no patient records. The written model was compared with the fitted model on draws of 100 to 5,000 development patients. The loop also drafted revisions to the rules that match a claim to a column and kept a revision only if more claims compiled. The Appendix gives the full procedure.

Ablation: no RSI loop. For each MIMIC-IV seed, every feature built in the first iteration was added to the baseline at once, with no gate and no further iteration, and the table was scored once on the holdout exactly as a loop is.

Comparator: general-purpose LLM. Qwen3-235B-A22B-Instruct-2507 [21] received the list of MIMIC-IV variable names with units and coverage and the outcome definition, and no patient data, abstracts or description of the loop. The instruction was: "You are given the list of variables available in a heart-failure patient dataset and the prediction target. Choose the five variables or simple derived variables (a cutoff, a ratio, or a most-recent value) that you expect to give the best prediction of death within 365 days of the first heart-failure diagnosis. Return only the list, one per line, each with a one-line definition that can be computed from the listed variables." It was called ten times, once per seed, under default sampling. Each returned list was built with the loop's feature-building rules, dropping variables that could not be built, and scored once on the holdout.

TABLE I

PATIENT CHARACTERISTICS OF THE THREE COHORTS. VALUES ARE N (%) UNLESS STATED; THE SYNTHETIC COLUMN DESCRIBES ALL 10,000 GENERATED RECORDS, AND ITS DEATHS ARE COUNTED AMONG THE 8,369 RECORDS WITH A KNOWN OUTCOME.

	Synthetic	MIMIC-IV	Health system
Patients, n	10,000	31,427	26,168
Age, years, median [IQR]	76 [64–85]	75 [65–84]	75 [64–84]
Female sex	4,969 (49.7)	14,704 (46.8)	12,649 (48.3)
Race			
White	4,094 (40.9)	22,470 (71.5)	10,499 (40.1)
Black or African American	2,348 (23.5)	3,583 (11.4)	6,324 (24.2)
Asian	373 (3.7)	668 (2.1)	930 (3.6)
Other or unknown	3,185 (31.9)	4,706 (15.0)	8,415 (32.2)
Ethnicity			
Hispanic or Latino	990 (9.9)	1,047 (3.3)	2,468 (9.4)
Not Hispanic or Latino	7,674 (76.7)	27,748 (88.3)	20,374 (77.9)
Unknown	1,336 (13.4)	2,632 (8.4)	3,326 (12.7)
Died within 365 days	1,514 (18.1)	8,282 (26.4)	4,798 (18.3)

III. RESULTS

Table I summarizes the three cohorts, and Table II presents the holdout results of the four experiments. Appendix Table A5 presents the ablation and the comparator.

Experiment 1: synthetic cohort. Across 15 runs, 450 proposals covered 84 concepts; 27 had eligible evidence and 23 reached evaluation (Table II). Only estimated GFR was accepted, in two runs, with AUROC gains of +0.013 on development folds and +0.010 on the holdout (95% CI -0.003 to $+0.021$). One supporting sentence reported 12% higher mortality per 10-unit decrease in eGFR [22]. Longer iteration budgets yielded no additional features. Most retrieved abstracts did not report a mortality effect for the proposed concept, so few candidates reached the gate. Development AUROC was 0.642 for demographics and diagnoses versus 0.685 for the cleaned table. Adding 52 literature-derived features reached 0.679, recovering 85% of the AUROC gap and 75% of the log-loss gap. Adding all 114 raw labs also reached 0.679, recovering 86% and 40%, respectively; shuffled literature features scored 0.631. Holdout AUROC gains over the minimal

TABLE II

HOLDOUT AUROC GAIN OVER THE BASELINE IN THE FOUR EXPERIMENTS. EXPERIMENTS 1 TO 3 GIVE THE MEAN ± SD ACROSS ACCEPTING LOOPS, ONE LOOP PER SEED, WITH THE LOWEST 95% CI BOUND, AND EXPERIMENT 4 GIVES THE GAIN OVER THE CORRECTED BASELINE [95% CI].

Experiment and arm	Iterations	Loops with ≥1 accepted feature	Features, mean	Holdout gain, logistic		Holdout gain, boosting
				mean ± SD	lowest CI bound	
Experiment 1: synthetic cohort						
RSI loop	1	1/5	–	+0.010	−0.003	–
RSI loop	3	0/5	–	–	–	–
RSI loop	5	1/5	–	+0.010	−0.003	–
Experiment 2: MIMIC-IV (baseline AUROC 0.747)						
RSI loop	1	9/10	0.9	+0.0038 ± 0.0015	−0.0005	+0.0031 ± 0.0014
RSI loop	3	10/10	2.4	+0.0065 ± 0.0021	+0.0011	+0.0045 ± 0.0024
RSI loop	5	10/10	4.2	+0.0103 ± 0.0025	+0.0027	+0.0075 ± 0.0027
Experiment 3: health-system cohort (baseline AUROC 0.708)						
RSI loop	1	10/10	1.0	+0.0176 ± 0.0053	+0.0087	+0.0066 ± 0.0026
RSI loop	3	10/10	3.0	+0.0363 ± 0.0080	+0.0147	+0.0121 ± 0.0023
RSI loop	5	10/10	4.9	+0.0469 ± 0.0045	+0.0268	+0.0139 ± 0.0021
Experiment 4: MIMIC-IV, second holdout (corrected baseline AUROC 0.747)						
RSI loop of Experiment 2	5	–	4.8	+0.0082 [+0.0037, +0.0128]		+0.0066
Corrected RSI loop	5	–	5	+0.0113 [+0.0068, +0.0156]		+0.0068
Lasso selection of laboratory tests	–	–	4.8	+0.0247 [+0.0180, +0.0316]		+0.0152
RSI loop, weights from published effect sizes	5	–	3	+0.0030 [+0.0011, +0.0051]		–
Weights written by a frontier model, no patient records	–	–	5	−0.0111 [−0.0186, −0.0045]		–
Weights written by a frontier model, used as a prior	–	–	5	+0.0118 [+0.0072, +0.0163]		–

baseline were +0.024 for literature features (95% CI +0.001 to +0.047) and +0.051 for the cleaned table.

Experiment 2: MIMIC-IV. Twenty-nine of 30 loops accepted at least one feature. The mean holdout AUROC gain was +0.004, +0.007 and +0.010 at 1, 3 and 5 iterations (Table II; baseline holdout AUROC 0.747), close to the corresponding development gains (+0.003, +0.005 and +0.009). At five iterations, every logistic and every boosting holdout CI excluded zero; at one iteration, 8 of 9 logistic CIs did. The most frequently accepted concepts were transferrin (13 of 30 loops), body-mass index (9), bilirubin (8), C-reactive protein (8) and total cholesterol (8), followed by HDL cholesterol (6), ferritin (6) and troponin T (5). Estimated GFR was proposed and built but not accepted. In the laboratory-recovery analysis, the literature-derived features recovered 64% of the holdout gain of the full laboratory panel over a table of demographics and diagnoses (Appendix Table A4).

Experiment 3: health-system cohort. All 30 loops accepted at least one feature, and every logistic holdout gain had a 95% CI that excluded zero (Table II; baseline holdout AUROC 0.708). The mean gain increased from +0.018 at one iteration to +0.047 at five, close to the corresponding development gains (+0.017, +0.035 and +0.045). Extraction in this cohort used the 7-billion-parameter open model running inside the institution; its features passed the same gate and gave the largest gains of the three cohorts. Estimated GFR was accepted in 29 of 30 loops, followed by alanine aminotransferase (12 loops), natriuretic peptides (11), anticoagulants (9), renal insufficiency (9) and potassium (7). Gains under the boosting model were smaller (+0.007 to +0.014), and 29 of 30 boosting CIs excluded zero. In the laboratory-recovery analysis, the literature-derived features recovered 86% of the holdout gain of the full laboratory panel (Appendix Table A3).

Experiment 4: RSI loop with written weights. On the second holdout of 6,286 patients (1,657 deaths), the loop of Experiment 2 gained +0.013 AUROC over the baseline as it was (0.743) and +0.008 over the corrected baseline (0.747), so about a third of its gain came from recorded-value indicators that were missing from the original baseline (Table II). The corrected loop accepted five features and gained +0.011 (95% CI +0.007 to +0.016), and lasso selection gained +0.025 (95% CI +0.018 to +0.032). Of 186 eligible claims, 18 (9.7%) compiled into a written weight; with these weights held fixed, the loop accepted three features and gained +0.003 (95% CI +0.001 to +0.005). The model whose 74 weights were written by the frontier model reached a holdout AUROC of 0.736 with no patient records, against 0.758 for the corrected loop's fitted model. With 1,000 or fewer local patients, the written weights with a locally fitted intercept gave lower log-loss than the fitted model. Self-revision of the loop's rules added no accepted feature. The Appendix gives the full results (Appendix Tables A6 to A12 and Fig. A1).

Ablation. Appendix Table A5 presents the ablation and the comparator next to the loop's five-iteration result; all three gains are over the same MIMIC-IV baseline. Without the loop, each seed's first-iteration candidates added at once (4.7 features on average; most often eGFR, angiotensin receptor antagonists and urea) gained +0.004, compared with +0.010 for the loop's 4.2 gated features, and the lowest CI bound fell below zero.

Comparator. The general-purpose LLM returned five variables that could be built in every call (0 of 50 dropped), most often age, NT-proBNP, creatinine and hemoglobin, all of which the baseline already held. Its features gained +0.003,

compared with +0.010 for the loop, and the lowest CI bound fell below zero.

IV. Discussion

Across the four experiments, five iterations of the RSI loop raised holdout AUROC by +0.008 to +0.047 over cleaned baselines that already held the standard predictors.

The features that the loop accepted are ones a cardiologist would recognize [10], [11], [23]. In the health-system cohort, the loop accepted estimated glomerular filtration rate in 29 of 30 loops, followed by alanine aminotransferase and natriuretic peptides. On MIMIC-IV, the loop most often accepted transferrin, body-mass index, bilirubin, C-reactive protein and total cholesterol. Most of these features mark renal function, cardiac stress, iron status, liver function or inflammation. Two of these already guide care: natriuretic peptides are recommended for risk stratification, and iron deficiency in heart failure with reduced ejection fraction can be treated with intravenous iron [23]. Each accepted feature is stored with the abstract sentence it came from, the measurement, the cutoff and the direction of effect. The eGFR feature accepted in the synthetic cohort, for example, carries the sentence reporting 12% higher mortality per 10-unit decrease in eGFR [22]. A clinician can therefore check a feature against its source before the feature enters a model, rather than trusting a column name.

These results bear on three hospital uses: risk stratification after a first heart-failure diagnosis, targeting of guideline-directed therapy, and timing of referral for advanced therapies or palliative care [23]. The need is large. In the three cohorts, 18% to 26% of patients died within a year of diagnosis (Table I). In a United States registry of outpatients with reduced ejection fraction, 1% of those eligible received target doses of all recommended drug classes [24]. Better ranking of patients by risk supports each of these uses, and the loop raised holdout AUROC from 0.708 to 0.755 in the health-system cohort and from 0.747 to 0.757 on MIMIC-IV. In the health-system cohort, the literature-derived features recovered 86% of the gain that a full laboratory panel adds to a table of demographics and comorbidities, so the literature pointed the loop to most of the prognostic value in the laboratory record. Beyond cardiology, the variable-preparation step that the loop automates precedes every clinical model built from structured records: data preparation takes 39% to 45% of analyst time [1], and the 30 health-system loops completed on one GPU in about 13 hours; these are illustrative rather than measured savings.

The same published evidence can also set a model's weights, which matters for a hospital with too few heart-failure patients to fit reliable weights of its own [25]. The loop already extracts an effect size with each claim, and weight-space programming lets a language model write the weights directly. In Experiment 4, a model whose weights were written with no patient records reached an AUROC of 0.736, and with a few hundred local patients the written weights gave lower log-loss than fitting, including fitting tuned for small samples; with thousands of patients, fitting was better. Starting the fit from the frontier model's written weights was never worse than fitting alone at any size tested, which agrees with earlier work in which published models and coefficients helped most in small samples [26], [27], [28]. Weights compiled claim by claim from abstracts were less useful: one eligible claim in ten compiled, and the three accepted weights were weaker than the local ones.

Several limitations should be considered. Evaluation is limited to one synthetic cohort, one health system and MIMIC-IV. Experiment 4 used MIMIC-IV only, so the health-system gains remain measured against a baseline without recorded-value indicators. Journal coverage was restricted and PubMed retrieval varies. The gate is heuristic and carries no error-rate guarantee [29]. The health-system baseline held no laboratory values, so its gains are over demographics, vital signs and comorbidities. All loops at a site share one holdout, and the holdout was scored twice for one MIMIC-IV loop because of an execution error, with identical readings and no selection on it. The second holdout of Experiment 4 is a new split of the same patients. Index-admission diagnosis codes may carry information from later in the stay. The concept pool bounds what the loop can propose. Only abstracts were read, and written weights could not be compiled for derived measures such as eGFR. Extraction varies between runs, so acceptance frequency across loops is the more reliable readout. In the health-system cohort, 41.6% of patients were absent from a truncated diagnosis extract, so comorbidity rates are lower bounds. The lasso comparator kept recorded-value indicators that the corrected loop could not accept, so that comparison favors lasso. The frontier model has read the medical literature, which includes studies of MIMIC, and the death count that the frontier model received covers the holdout patients, so the written weights are not independent of this database. Future work should evaluate the loop on larger, multi-institutional, temporally separated cohorts and repeat Experiment 4 in a second health system, including hospitals with few heart-failure patients. The loop can also read full-text articles, build derived measures and composite risk scores, and be compared with a language model that is shown the baseline. Clinicians should review the accepted features and written weights, and a prospective study should test whether the resulting models change therapy or referral decisions. Until then, nMAS should remain a review-support workflow rather than a replacement for clinical judgment.

References

[1] Anaconda, Inc., "The state of data science 2020," https://www.anaconda.com/resources/whitepaper/state-of-data-science-2020, 2020.

[2] G. C. Fonarow *et al.*, "HF STATS 2025: Heart failure epidemiology and outcomes statistics," *Journal of Cardiac Failure*, vol. 32, no. 2, pp. 439–498, 2026.

[3] G. L. Di Tanna, H. Wirtz, K. L. Burrows, and G. Globe, "Evaluating risk prediction models for adults with heart failure: a systematic literature review," *PLoS ONE*, vol. 15, no. 1, p. e0224135, 2020.

[4] N. Hollmann, S. Müller, and F. Hutter, "Large language models for automated data science: Introducing CAAFE for context-aware automated feature engineering," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.

[5] S. R. Shimgekar, S. Vassef, A. Goyal, N. Kumar, and K. Saha, "Agentic AI framework for end-to-end medical data inference," arXiv:2507.18115, 2025.

[6] S. R. Shimgekar *et al.*, "Tracing the heart: An evidence-linked pipeline for heart-failure feature engineering," arXiv:2608.06366, 2026.

[7] A. Novikov *et al.*, "AlphaEvolve: A coding agent for scientific and algorithmic discovery," arXiv:2506.13131, 2025.

[8] J. Zhang, S. Hu, C. Lu, R. Lange, and J. Clune, "Darwin Gödel Machine: Open-ended evolution of self-improving agents," arXiv:2505.22954, 2025.

[9] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.

[10] W. C. Levy *et al.*, "The Seattle Heart Failure Model: Prediction of survival in heart failure," *Circulation*, vol. 113, no. 11, pp. 1424–1433, 2006.

[11] S. J. Pocock, C. A. Ariti, J. J. V. McMurray, A. Maggioni, L. Køber, I. B. Squire, K. Swedberg, J. Dobson, K. K. Poppe, G. A. Whalley, and R. N. Doughty, "Predicting survival in heart failure: a risk score based on 39,372 patients from 30 studies," *European Heart Journal*, vol. 34, no. 19, pp. 1404–1413, 2013.

[12] I. C. Wiest *et al.*, "Privacy-preserving large language models for structured medical information retrieval," *npj Digital Medicine*, vol. 7, p. 257, 2024.

[13] DeepSeek-AI, "DeepSeek-V4: Towards highly efficient million-token context intelligence," Technical report, 2026.

[14] R. Takehi, J. Teoh, A. Zhang, and O. Khattab, "Sparks of weight-space programming: LLMs can natively read and write neural network weights," MIT CSAIL blog, https://rikiyat.github.io/blog/weight-space-programming/, 2026.

[15] J. Hu, C. Tang, W. Li, H. Zhu, C.-M. Feng, Q. Zhang, Y. Chen, and M. Hu, "Exploring weight-space programming: capabilities, generalization, and limitations," arXiv:2611.09382, 2026.

[16] E. Klang *et al.*, "A strategy for cost-effective large language model use at health system-scale," *npj Digital Medicine*, vol. 7, p. 320, 2024.

[17] A. E. W. Johnson, L. Bulgarelli, L. Shen *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, 2023.

[18] A. Johnson *et al.*, "MIMIC-IV (version 3.1)," PhysioNet, doi:10.13026/kpb9-mt58, 2024.

[19] T. Pollard, B. E. Moody, L. Lehman, B. Gow, C. Fernandes, C. Xie, A. Johnson, R. G. Mark, and T. Heldt, "PhysioNet as a global platform for biomedical research," *Nature Health*, 2026.

[20] Qwen Team, "Qwen2.5 technical report," arXiv:2412.15115, 2024.

[21] A. Yang *et al.*, "Qwen3 technical report," arXiv:2505.09388, 2025.

[22] D. Kotecha *et al.*, "Impact of renal impairment on beta-blocker efficacy in patients with heart failure," *Journal of the American College of Cardiology*, vol. 74, no. 23, pp. 2893–2904, 2019, PMID: 31806133.

[23] P. A. Heidenreich *et al.*, "2022 AHA/ACC/HFSA guideline for the management of heart failure," *Circulation*, vol. 145, no. 18, pp. e895–e1032, 2022.

[24] S. J. Greene *et al.*, "Medical therapy for heart failure with reduced ejection fraction: the CHAMP-HF registry," *Journal of the American College of Cardiology*, vol. 72, no. 4, pp. 351–366, 2018.

[25] R. D. Riley, K. I. E. Snell, G. P. Martin, R. Whittle, L. Archer, M. Sperrin, and G. S. Collins, "Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small," *Journal of Clinical Epidemiology*, vol. 132, pp. 88–96, 2021.

[26] E. W. Steyerberg, G. J. J. M. Borsboom, H. C. van Houwelingen, M. J. C. Eijkemans, and J. D. F. Habbema, "Validation and updating of predictive logistic regression models: a study on sample size and shrinkage," *Statistics in Medicine*, vol. 23, no. 16, pp. 2567–2586, 2004.

[27] T. P. A. Debray, H. Koffijberg, D. Lu, Y. Vergouwe, E. W. Steyerberg, and K. G. M. Moons, "Incorporating published univariable associations in diagnostic and prognostic modeling," *BMC Medical Research Methodology*, vol. 12, p. 121, 2012.

[28] G. P. Martin, M. A. Mamas, N. Peek, I. Buchan, and M. Sperrin, "A multiple-model generalisation of updating clinical prediction models," *Statistics in Medicine*, vol. 37, no. 8, pp. 1343–1358, 2018.

[29] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[30] J. G. Ibrahim, M.-H. Chen, Y. Gwon, and F. Chen, "The power prior: theory and applications," *Statistics in Medicine*, vol. 34, no. 28, pp. 3724–3749, 2015.

[31] A. Capstick, R. G. Krishnan, and P. Barnaghi, "AutoElicit: using large language models for expert prior elicitation in predictive modelling," arXiv:2411.17284, 2024.

APPENDIX

Overview. The Appendix has four parts. Tables A1 and A2 give the patient flow and the full characteristics of the health-system cohort. Tables A3 and A4 give the laboratory-recovery analysis in the health-system cohort and on MIMIC-IV. Table A5 and the paragraph above it give the ablation and the comparator. The remaining pages give the additional methodology and results of Experiment 4, with Tables A6 to A12 and Fig. A1; the log-loss table and figure sit beside the paragraph that reports them.

TABLE A1
PATIENT FLOW IN THE HEALTH-SYSTEM COHORT.

Step	Patients
First heart-failure diagnosis in the export	33,914
Excluded: alive with <365 days of follow-up	4,331
Excluded: unknown vital status	2,570
Excluded: deceased without a death date	823
Excluded: no index date, or death before index	22
Analyzed (known 365-day outcome)	26,168
Development set (80%)	20,934 (3,838 deaths)
Reserved holdout (20%)	5,234 (960 deaths)

TABLE A2
FULL CHARACTERISTICS OF THE HEALTH-SYSTEM COHORT BY 365-DAY OUTCOME; N (%) UNLESS STATED, MEDIANS [IQR] AMONG PATIENTS WITH A RECORDED VALUE.

Characteristic	Survivors at 365 days	Deaths within 365 days
Patients, n	21,370	4,798
Age, years, median [IQR]	74 [63–84]	79 [70–87]
Age group: under 65	6,010 (28.1)	784 (16.3)
65–74	4,801 (22.5)	968 (20.2)
75–84	5,668 (26.5)	1,523 (31.7)
85 or older	4,877 (22.8)	1,522 (31.7)
Female sex	10,389 (48.6)	2,260 (47.1)
Race: White	8,686 (40.6)	1,813 (37.8)
Black or African American	5,469 (25.6)	855 (17.8)
Asian	768 (3.6)	162 (3.4)
Hispanic (recorded as race)	88 (0.4)	14 (0.3)
Other	5,706 (26.7)	1,635 (34.1)
Unknown, declined or missing	653 (3.1)	319 (6.6)
Ethnicity: not Hispanic or Latino	16,695 (78.1)	3,679 (76.7)
Hispanic or Latino	2,195 (10.3)	273 (5.7)
Unknown, declined or other	2,480 (11.6)	846 (17.6)
BMI recorded	7,881 (36.9)	2,528 (52.7)
BMI, kg/m^2, median [IQR]	28.3 [24.2–33.7]	26.2 [22.5–31.1]
Systolic BP recorded	8,035 (37.6)	2,573 (53.6)
Systolic BP, mmHg, median [IQR]	124 [111–140]	115 [101–132]
Diastolic BP, mmHg, median [IQR]	68 [60–76]	63 [55–72]
Pulse, beats/min, median [IQR]	76 [67–87]	77 [65–91]
Comorbidities: chronic kidney disease	6,917 (32.4)	1,589 (33.1)
Atrial fibrillation	7,174 (33.6)	1,637 (34.1)
Coronary artery disease	8,790 (41.1)	1,746 (36.4)
Cardiomyopathy	4,249 (19.9)	736 (15.3)
Diabetes	7,199 (33.7)	1,340 (27.9)
COPD	4,847 (22.7)	994 (20.7)
Obesity	5,326 (24.9)	837 (17.4)
Pulmonary hypertension	3,814 (17.8)	910 (19.0)
Sepsis	3,282 (15.4)	1,206 (25.1)
Anemia	7,472 (35.0)	1,872 (39.0)
Present in diagnosis extract	12,879 (60.3)	2,391 (49.8)

Outcome: death within 365 days of the first heart-failure diagnosis; day-0 deaths count as deaths. Vital signs are a single undated value per patient, recorded for about 40%. Comorbidities are taken from a truncated diagnosis extract, so 41.6% of patients do not appear in it and have every flag set to 0; prevalences are lower bounds.

TABLE A3

LABORATORY-RECOVERY ANALYSIS, HEALTH-SYSTEM HOLDOUT (5,234 PATIENTS): AUROC OF THE LAB-FREE MINIMAL TABLE AND CHANGE RELATIVE TO IT [95% CI]; THE LITERATURE FEATURES RECOVERED 86% (LOGISTIC) AND 96% (BOOSTING) OF THE ALL-LABORATORY GAIN. THE LITERATURE ROWS POOL EVERY LITERATURE-DERIVED FEATURE BUILT BY THE 30 LOOPS, ACCEPTED OR NOT, WITHOUT THE GATE, SO THEIR GAINS ARE NOT COMPARABLE TO A SINGLE LOOP'S ACCEPTED SET.

Arm	Logistic	Boosting
Minimal (demographics, diagnoses)	0.680	0.692
+ literature (61 columns)	+0.0791 [+0.0649, +0.0931]	+0.0947 [+0.0782, +0.1109]
+ all labs (136 columns)	+0.0924 [+0.0774, +0.1081]	+0.0990 [+0.0823, +0.1153]
+ placebo	−0.0003 [−0.0042, +0.0039]	−0.0078 [−0.0140, −0.0013]
Cleaned baseline	+0.0275 [+0.0155, +0.0389]	+0.0779 [+0.0629, +0.0938]
Cleaned + literature	+0.0912 [+0.0754, +0.1062]	+0.1026 [+0.0853, +0.1202]

TABLE A4

LABORATORY-RECOVERY ANALYSIS, MIMIC-IV HOLDOUT (6,286 PATIENTS): AUROC OF THE MINIMAL TABLE OF DEMOGRAPHICS AND DIAGNOSES AND CHANGE RELATIVE TO IT [95% CI]; THE LITERATURE FEATURES RECOVERED 64% (LOGISTIC) AND 70% (BOOSTING) OF THE ALL-LABORATORY GAIN. THE LITERATURE ROWS POOL EVERY LITERATURE-DERIVED FEATURE BUILT BY THE 30 LOOPS, ACCEPTED OR NOT, WITHOUT THE GATE, SO THEIR GAINS ARE NOT COMPARABLE TO A SINGLE LOOP'S ACCEPTED SET.

Arm	Logistic	Boosting
Minimal (demographics, diagnoses)	0.711	0.730
+ literature (76 columns)	+0.0484 [+0.0393, +0.0579]	+0.0511 [+0.0419, +0.0604]
+ all labs (602 columns)	+0.0762 [+0.0638, +0.0878]	+0.0728 [+0.0614, +0.0847]
+ placebo	−0.0020 [−0.0046, +0.0009]	−0.0030 [−0.0065, +0.0003]
Cleaned baseline	+0.0362 [+0.0279, +0.0444]	+0.0358 [+0.0276, +0.0450]
Cleaned + literature	+0.0555 [+0.0460, +0.0653]	+0.0548 [+0.0461, +0.0648]

Ablation and comparator: additional results. Table A5 lists the ablation and the comparator next to the RSI loop at five iterations, with ten seeds each. The 95% CI of the logistic gain lay above zero in all ten seeds for the loop, in seven for the ablation and in two for the comparator. Without the loop, the concepts built in the first iteration were most often estimated GFR (10 of 10 seeds), angiotensin receptor antagonists (8), urea (5), transferrin (3) and stroke (3). The general-purpose LLM most often chose age (10 of 10 calls), NT-proBNP (10), creatinine (9), hemoglobin (9), sodium (4) and estimated GFR (3). Of the 50 variables that the general-purpose LLM chose, 42 were variables that the baseline already held.

TABLE A5

ABLATION AND COMPARATOR ON MIMIC-IV, TEN SEEDS EACH, NEXT TO THE RSI LOOP AT FIVE ITERATIONS: HOLDOUT AUROC GAIN OVER THE SAME BASELINE, MEAN ± SD AND RANGE ACROSS SEEDS, THE LOWEST 95% CI BOUND, AND THE SEEDS WHOSE CI LAY ABOVE ZERO.

Arm	Features added, mean	Holdout gain, logistic mean ± SD	Range across seeds	Lowest CI bound	Seeds with CI above zero	Holdout gain, boosting
RSI loop	4.2	+0.0103 ± 0.0025	+0.0063 to +0.0144	+0.0027	10/10	+0.0075 ± 0.0027
No RSI loop	4.7	+0.0035 ± 0.0022	−0.0004 to +0.0074	−0.0010	7/10	+0.0027 ± 0.0015
General-purpose LLM	5.0	+0.0029 ± 0.0036	+0.0009 to +0.0121	−0.0008	2/10	+0.0014 ± 0.0028

Experiment 4: Additional methodology. In Experiments 1 to 3 the loop changed its feature table, and every weight in the model was fitted to patients. Experiment 4 applied the same cycle of proposing, testing and keeping to two further objects: the weights of the model and the loop's own rules. The experiment ran on MIMIC-IV in four parts, and the test for each part was fixed before that part was run.

Corrected loop. Five corrections were made before any weight was written. Baseline columns with missing values received an indicator that a value was recorded, as the loop's features already had; this was the corrected baseline. Each literature feature entered the gate as its value only, and its indicator was scored separately and never accepted. Each concept was bound to a single laboratory test. Coverage counted patients with a usable value. Every eligible concept was scored in each iteration rather than a sample of ten. Patients were split into development (80%) and holdout (20%) sets with a new seed, because the first holdout had been scored in the earlier experiments, and every result of Experiment 4 used this second holdout. The loop of Experiment 2 was rerun on the new split with ten seeds. The corrected loop involved no random draw and was scored once. Both ran five iterations with Qwen2.5-7B-Instruct and read one shared set of extracted claims. As a comparator, a lasso logistic regression selected laboratory tests from all those recorded for at least 500 development patients. For each seed it was limited to the number of features that the loop of Experiment 2 accepted, and each selected test entered with its recorded-value indicator.

Weights from published effect sizes. Each extracted claim that reported an odds, hazard or risk ratio for death was compiled into a weight, the natural logarithm of the ratio, for the cutoff or the presence that the claim named. Four checks had to pass: the number appeared in the quoted sentence, its sign agreed with the stated direction, the unit converted to that of the column, and the measurement matched the column. Claims for the same feature were pooled. The corrected loop then ran with each candidate's weight held at its written value. In a second form the written weight was the center of a normal prior [30] whose variance was the squared standard error from the reported confidence interval plus an allowance of 0.15^2 for disagreement between studies. For the prior form, three further checks refused a cutoff absent from the quoted sentence, a comparison between treatments and an outcome other than death, and a claim from another patient group entered at half precision.

Three tests followed. Replication compared each written weight with the weight fitted on the development set. The small-site test asked whether written weights helped when local patients were few. Patients were divided by the three-year period to which MIMIC-IV assigns each patient. The loop ran on 2008 to 2013, the period 2014 to 2016 was left unused, and models were refitted on draws of 250, 500, 1,000 and 2,000 patients from one half of 2017 to 2019 and scored on the other half, with 20 draws per size. The test was met if the prior form had lower log-loss than fitted weights at 250, 500 or 1,000 patients, with a 98.33% interval over draws (95% corrected for three sizes) excluding zero. A first form of this test had used fixed weights and AUROC on the second holdout; the form with the prior and log-loss was specified after the first had been read. The transfer test compared the written weights with weights fitted on the 2008 to 2013 patients, both held fixed in the later period.

Self-revision of the loop's rules. The loop proposed changes to the rules that match a claim to a column, which cover names, abbreviations and units. In each generation a language model received the log of failed claims with their reasons and the hospital's lists of laboratory, medication and diagnosis names, and no patient data, and drafted rule revisions. A critic, a fixed set of checks in code, rejected drafts that were malformed, duplicated, named a paper or a claim, or used a name outside the lists. Each remaining draft was replayed on the claims and kept only if at least one more claim compiled and none was lost, up to five per generation. The next generation started from the revised rules, for up to three generations, and a chain that kept drafts by replay stopped when a generation kept nothing. Revisions were drafted on the claims of 100 papers; the claims of 43 other papers were held back and used only to measure whether the revisions carried over. Two arms were run first: Qwen2.5-72B-Instruct [20] as drafter, and a control in which five passing drafts of the same drafter were kept at random. A third arm, with Claude Fable 5.2 (Anthropic) as drafter, was added after the first two had been read, to test whether the drafter had limited the result. Self-revision was counted as supported only if, from the starting rules to the last generation, more held-back claims compiled and the gate accepted more features. The starting rules had been adjusted partly on the held-back claims, and the answer key for those claims came from a language model rather than a clinician.

Weights written by a frontier model. A frontier model, Claude Fable 5.2, wrote every weight of the corrected loop's final model. This follows weight-space programming, defined in the Methods, and earlier work in which language models were asked to supply priors [31]. Claude Fable 5.2 received a description of each column (name, meaning, unit, coding, timing and handling of missing values), the outcome definition, the published size and death count of the cohort, and the quoted claims behind the loop's accepted features. No patient record was sent, and no statistic of the cohort other than that published size and death count. In three independent sessions the frontier model returned an intercept and, for each column, a reference value, a weight per unit and a standard deviation expressing its uncertainty; the median of the three was used. The sessions were instructed to read only the material they were given but were not technically prevented from reading more. The written model was used in four ways: as written, with no patient records; with only the intercept fitted to local patients; with the intercept and one overall slope fitted [26]; and with each written weight as the center of a normal prior with the stated standard deviation. These were compared with the loop's fitted logistic model on draws of 100, 250, 500, 1,000 and 5,000 development patients, 20 draws per size,

and on all development patients, each scored on the second holdout. The test was met if the written weights with a local intercept had lower log-loss than the fitted model at 100, 250, 500 or 1,000 patients, with a 98.75% interval over draws (95% corrected for four sizes) excluding zero. Because the loop's logistic model was not tuned for small samples, the comparison was repeated on a split of the development set against ridge and lasso logistic models whose penalty was chosen by five-fold cross-validation within each draw.

Experiment 4: Additional results. Table A6 presents each model of Experiment 4 on the second holdout of 6,286 patients (1,657 deaths), and Table A7 and Fig. A1 the comparison by number of local patients.

Corrected loop. On the second holdout, the loop of Experiment 2 gained +0.013 AUROC over the baseline as it was (0.743) and +0.008 over the corrected baseline (0.747). About a third of its gain therefore came from recorded-value indicators that the baseline had lacked. On the development folds of the first split, 53% of the five-iteration gain of Experiment 2 remained when features entered as their value only against the corrected baseline. The corrected loop accepted five features (cardiogenic shock, bilirubin, chloride, transferrin and HDL cholesterol) and gained +0.011 over the corrected baseline (95% CI +0.007 to +0.016). Lasso selection, limited to the number of features that the loop of Experiment 2 accepted (4.8 on average), gained +0.025 (95% CI +0.018 to +0.032). Under the boosting model the gains were +0.007 for both loops and +0.015 for lasso.

Weights from published effect sizes. Of 186 eligible claims from 143 papers, 18 (9.7%) compiled into a written weight on MIMIC-IV, giving 13 written features after pooling (Appendix Table A8). The most common reason for failure was a concept with no column in the cohort (63 claims). Every written feature was a cutoff or the presence of a diagnosis or drug; no claim stated per unit of a measurement compiled. With the written weights held fixed, the loop accepted three features: cardiogenic shock, ferritin below 100 ng/mL and cachexia. They gained +0.003 (95% CI +0.001 to +0.005), against +0.004 for the prior form and +0.004 for the same three features with fitted weights (Table A6). On the synthetic cohort 16 claims compiled and the loop accepted no written feature. In replication, all three accepted written weights were smaller in size than the fitted weights: 0.65 against 0.90 for cardiogenic shock, -0.09 against -0.43 for low ferritin and 0.21 against 0.72 for cachexia (Appendix Table A9). Of the seven rejected candidates, four agreed with the fitted weight, two were larger in size than the fitted weight and one was reversed. Several weights compiled from abstracts rested on a claim from a different population, such as a pediatric cohort behind the weight for cardiogenic shock. In the small-site test, the prior form had lower log-loss than fitted weights at 250 patients (-0.013) and at 500 patients (-0.002); both 98.33% intervals excluded zero, and the test was met. There was no difference at 1,000 patients, and fitted weights were slightly better at 2,000 (+0.001; Appendix Table A10). On the patients of 2020 to 2022, a second test set with no pre-stated rule, the prior form

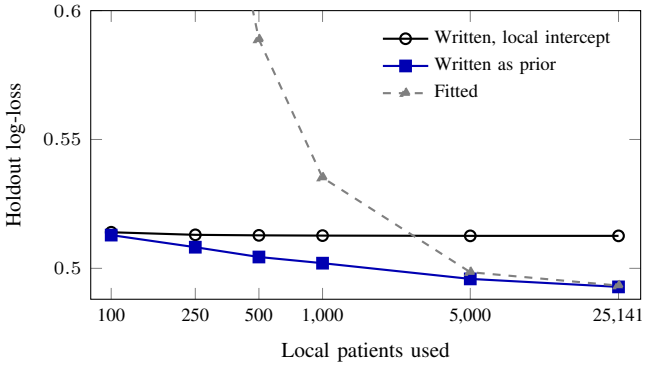


Fig. A1. Holdout log-loss by number of local patients, for weights written by the frontier model and for the loop's fitted model. The fitted model's values at 100 and 250 patients (0.849 and 0.706) lie above the plotted range.

was better at every size. The fixed and prior forms differed by 0.001 or less at every size up to 2,000 patients, so the benefit came from using an outside number rather than from letting it move. The first form of the test, with fixed weights scored by AUROC on the second holdout, showed no advantage at any size. The transfer test was not met: held fixed in the later period, the written weights had higher log-loss than the weights fitted on earlier patients (+0.002, 95% CI +0.000 to +0.003).

Self-revision. With the 72-billion-parameter drafter, 7 of 54 first-generation drafts passed the critic and 2 were kept; no second-generation draft passed, and the chain stopped. With Claude Fable 5.2, 25 of 45 drafts passed in the first generation and 9 of 26 in the second; 4 and 5 were kept, and none in the third (Appendix Table A11). Its revisions raised the claims that compiled on the drafting papers from 10 to 17 of 125 and on the held-back papers from 4 to 5 of 61. The corresponding counts were 10 to 13 and 4 to 4 for the 72-billion-parameter drafter, and 10 to 12 and 4 to 4 for the random control. The gate accepted the same three features in every generation of every arm, so self-revision was not supported by the test. Read against their source claims, three of the five revisions kept in the second generation of Claude Fable 5.2 built a column that read wrong or matched the claim only loosely. Low cardiac output after cardiac surgery was matched to diagnoses of cardiogenic shock, and right ventricular dysfunction, an echocardiographic finding, was matched to diagnoses of right heart failure. The gate accepted none of the features that the revisions produced. The column built for low cardiac output passed the gate's thresholds on its own and was rejected only because it marked the same patients as the cardiogenic-shock feature already accepted.

Weights written by a frontier model. The model held 74 columns: 69 baseline columns and the five accepted features. The three sessions never gave weights of opposite sign and agreed closely, so they give one written model rather than three. The median weight had the sign of the fitted weight for 53 of 74 columns, and for 65 of 74 the fitted weight lay within two of the stated standard deviations. As writ-

TABLE A6

EXPERIMENT 4 ON THE SECOND MIMIC-IV HOLDOUT, LOGISTIC MODEL: AUROC, GAIN OVER THE CORRECTED BASELINE [95% CI] AND LOG-LOSS. ROWS WITH 4.8 FEATURES ARE MEANS OVER TEN SEEDS; THE LASSO LABORATORY TESTS KEPT THEIR RECORDED-VALUE INDICATORS.

Model	Source of the weights	Features	AUROC	Gain over corrected baseline	Log-loss
Corrected baseline	fitted to patients	–	0.747	–	0.500
RSI loop of Experiment 2	fitted to patients	4.8	0.755	+0.0082 [+0.0037, +0.0128]	0.495
Corrected RSI loop	fitted to patients	5	0.758	+0.0113 [+0.0068, +0.0156]	0.493
Lasso selection of laboratory tests	fitted to patients	4.8	0.772	+0.0247 [+0.0180, +0.0316]	0.484
RSI loop with written weights	published effect sizes, held fixed	3	0.750	+0.0030 [+0.0011, +0.0051]	0.498
RSI loop with written weights	published effect sizes, as a prior	3	0.751	+0.0043 [+0.0012, +0.0075]	0.497
The same three features	fitted to patients	3	0.751	+0.0044 [+0.0011, +0.0078]	0.497
Corrected loop's model, every weight written	frontier model, no patient records	5	0.736	−0.0111 [−0.0186, −0.0045]	0.513
Corrected loop's model, every weight written	frontier model, as a prior	5	0.759	+0.0118 [+0.0072, +0.0163]	0.493

TABLE A7

LOG-LOSS ON THE SECOND HOLDOUT BY NUMBER OF LOCAL PATIENTS, MEAN OF 20 DRAWS, WITH WEIGHTS WRITTEN BY THE FRONTIER MODEL. THE LAST COLUMN IS LOCAL INTERCEPT MINUS FITTED [98.75% INTERVAL OVER DRAWS].

Local patients	Written, no patient records	Written, local intercept	Written as prior	Fitted	Local intercept minus fitted
100	0.5126	0.5140	0.5129	0.8488	−0.3348 [−0.3688, −0.3015]
250	0.5126	0.5130	0.5082	0.7064	−0.1934 [−0.2208, −0.1682]
500	0.5126	0.5128	0.5044	0.5888	−0.0761 [−0.0845, −0.0675]
1,000	0.5126	0.5127	0.5020	0.5352	−0.0225 [−0.0276, −0.0176]
5,000	0.5126	0.5126	0.4959	0.4985	+0.0141 [+0.0131, +0.0149]
All 25,141	0.5126	0.5126	0.4928	0.4933	+0.0193 (one fit)

ten, with no patient records, the model reached a holdout AUROC of 0.736 and a log-loss of 0.513. Both were worse than the corrected baseline fitted on 25,141 patients (0.747 and 0.500) and the corrected loop's fitted model (0.758 and 0.493; Table A6). With fewer local patients the order reversed (Table A7, Fig. A1). The written weights with a local intercept had lower log-loss than the loop's fitted model at 100, 250, 500 and 1,000 patients in all 20 draws, and the test was met at each size; at 5,000 patients the fitted model was better. Fitting an overall slope as well as the intercept did not help below 1,000 patients (log-loss 0.521, 0.518 and 0.514 at 100, 250 and 500). Used as a prior, the written weights gave the lowest log-loss of all forms from 250 patients upward and equaled the fitted model with all patients (AUROC 0.759 against 0.758). Against ridge and lasso models tuned within each draw, the written weights with a local intercept were still better at 100, 250 and 500 patients in all 20 draws. They could not be separated from tuned lasso at 1,000 patients and were worse at 5,000 (Appendix Table A12). Tuning removed most of the margin: on the same draws at 100 patients, the difference in log-loss was 0.064 against tuned ridge and 0.340 against the loop's model. The prior form was better than both tuned models up to 5,000 patients and equal to them with all patients. The intervals in these small-sample comparisons reflect variation between draws of local patients on one test set rather than between test patients, and no nonlinear fitted model was compared.

TABLE A8

CLAIMS COMPILED INTO WRITTEN WEIGHTS IN EXPERIMENT 4, BY COHORT; THE 186 ELIGIBLE CLAIMS CAME FROM 143 PAPERS. REASONS ARE FOR THE FOUR-CHECK COMPILE.

	MIMIC-IV	Synthetic
Eligible claims	186	186
Matched to a column	65	55
Compiled into a weight, four checks	18 (9.7%)	16 (8.6%)
Written features after pooling	13	11
Compiled into a weight, seven checks	14 (7.5%)	12 (6.5%)
Written features after pooling	10	8
Most common reasons a claim did not compile		
No column for the concept	63	74
No column of the kind the claim names	18	25
Name given only as an abbreviation	12	9
Measurement differs from the column	10	10
Comparison of two active groups	10	10
Several ratios in one claim	9	6
Derived measure (ratio, index or score)	8	6
Words of the measurement left unmatched	8	8

TABLE A9

REPLICATION ON THE MIMIC-IV DEVELOPMENT SET: WRITTEN WEIGHT AND FITTED WEIGHT (LOG ODDS RATIO) [95% CI]. THE CLASS IS "AGREES" IF THE WRITTEN WEIGHT LIES INSIDE THE INTERVAL, "REVERSED" IF THE FITTED WEIGHT HAS THE OTHER SIGN AND ITS INTERVAL EXCLUDES ZERO, AND OTHERWISE "WEAKER" OR "STRONGER" BY SIZE. THE TROPONIN I WEIGHT WAS FITTED ON 32 PATIENTS WITH A VALUE.

Feature	Written	Fitted	Class	Accepted
Cardiogenic shock	$+0.6471$	$+0.8957$ [$+0.7701$, $+1.0182$]	weaker	yes
Ferritin $<$ 100 ng/mL	-0.0943	-0.4339 [-0.5555, -0.3282]	weaker	yes
Cachexia	$+0.2070$	$+0.7190$ [$+0.4635$, $+0.9403$]	weaker	yes
Amiodarone	$+0.5188$	$+0.3911$ [$+0.0958$, $+0.6942$]	agrees	no
Bundle-branch block	$+0.5072$	-0.1005 [-0.2235, $+0.0457$]	stronger	no
Digoxin $\geq$ 1.2 ng/mL	$+0.4447$	$+0.2045$ [-0.0673, $+0.5063$]	agrees	no
Hemoglobin $<$ 12 g/dL	-0.4005	$+0.0642$ [-0.0353, $+0.1866$]	stronger	no
Troponin I $>$ 0.09 ng/mL	$+1.1086$	-5.3140 [-5.7970, -3.7867]	reversed	no
Uric acid $>$ 7 mg/dL	$+0.1398$	$+0.0573$ [-0.1210, $+0.2359$]	agrees	no
Valsartan	-0.3711	-0.1949 [-0.4222, -0.0027]	agrees	no

TABLE A10

SMALL-SITE TEST ON MIMIC-IV DIVIDED BY ADMISSION PERIOD: LOG-LOSS OF THE PRIOR FORM MINUS FITTED WEIGHTS [98.33% INTERVAL OVER 20 DRAWS], AND THE DRAWS IN WHICH THE PRIOR FORM WAS BETTER. THE LOOP RAN ON 16,976 PATIENTS OF 2008 TO 2013; PATIENTS OF 2014 TO 2016 WERE NOT USED; LOCAL PATIENTS WERE DRAWN FROM 2,544 PATIENTS OF 2017 TO 2019 AND SCORED ON THE OTHER 2,544 AND ON 3,749 PATIENTS OF 2020 TO 2022. THE PRE-STATED TEST USED THE 2017 TO 2019 SET.

Local	Test set, 2017 to 2019		Test set, 2020 to 2022	
patients	Prior minus fitted	Draws	Prior minus fitted	Draws
250	-0.0132 [-0.0193, -0.0076]	17/20	-0.0190 [-0.0258, -0.0123]	19/20
500	-0.0021 [-0.0043, -0.0003]	13/20	-0.0052 [-0.0081, -0.0029]	20/20
1,000	$+0.0001$ [-0.0006, $+0.0009$]	8/20	-0.0017 [-0.0025, -0.0009]	18/20
2,000	$+0.0006$ [$+0.0005$, $+0.0007$]	0/20	-0.0006 [-0.0009, -0.0004]	18/20

TABLE A11

SELF-REVISION OF THE LOOP'S RULES ON MIMIC-IV, FROM THE STARTING RULES TO THE LAST GENERATION. THE ANSWER KEY MARKED 10 OF THE 61 HELD-BACK CLAIMS AS CLAIMS THAT SHOULD COMPILE.

	72B drafter	72B drafts, kept at random	Claude Fable 5.2 drafter
Generations run	2	3	3
Drafts passing the critic, generation 1	7 of 54	8 of 52	25 of 45
Revisions kept in all	2	10	9
Claims compiled, drafting papers (of 125)	10 to 13	10 to 12	10 to 17
Claims compiled, held-back papers (of 61)	4 to 4	4 to 4	4 to 5
of these, agreeing with the answer key	3 to 3	3 to 3	3 to 4
Written features	10 to 12	10 to 12	10 to 16
Features accepted by the gate	3 to 3	3 to 3	3 to 3

TABLE A12

COMPARISON WITH FITTED MODELS TUNED FOR SMALL SAMPLES, ON A SPLIT OF THE MIMIC-IV DEVELOPMENT SET (DRAWS FROM 20,112 PATIENTS, SCORED ON 5,029): LOG-LOSS, MEAN OF 20 DRAWS, AND LOCAL INTERCEPT MINUS EACH TUNED MODEL [98.75% INTERVAL OVER DRAWS].

Local patients	Written, local intercept	Written as prior	Loop's fitted model	Tuned ridge	Tuned lasso	Local intercept minus tuned ridge	Local intercept minus tuned lasso
100	0.4992	0.4988	0.8392	0.5631	0.5737	−0.0638 [−0.0900, −0.0482]	−0.0745 [−0.0960, −0.0601]
250	0.4990	0.4976	0.6892	0.5327	0.5400	−0.0337 [−0.0396, −0.0282]	−0.0410 [−0.0526, −0.0310]
500	0.4988	0.4959	0.5703	0.5168	0.5166	−0.0180 [−0.0216, −0.0144]	−0.0178 [−0.0244, −0.0121]
1,000	0.4985	0.4927	0.5199	0.5051	0.5004	−0.0066 [−0.0096, −0.0040]	−0.0018 [−0.0040, +0.0002]
5,000	0.4984	0.4879	0.4913	0.4895	0.4892	+0.0089 [+0.0083, +0.0095]	+0.0091 [+0.0085, +0.0098]
All 20,112	0.4983	0.4845	0.4845	0.4846	0.4845	+0.0138 (one fit)	+0.0138 (one fit)