LLM-Steered Clinical Knowledge Discovery for CVD Outcome Modeling

Yuxuan Liang[1 [0009−0000−8719−7472]], Xuanang Xu[1 [0000−0002−6045−8457]], Ge Wang[1 [0000−0002−2656−7705]], and Pingkun Yan[1 [0000−0002−9779−2141]⋆]

Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
yanp2@rpi.edu

Abstract. Cardiovascular disease (CVD) modeling has become highly accurate, yet most models remain "black boxes" that prioritize predictive scores over reliable and reusable clinical knowledge. We propose Evidence-Grounded Learning (EGL), which repeatedly tests candidate variable sets under a fixed evaluation budget and records auditable evidence about which factors and pairwise relationships are consistently supported. EGL–LLM adds an interpretable large language model (LLM) control plane that uses a semantic agenda to decide what to test next, while keeping the evidence definitions and budget unchanged. On the COPDGene cohort, EGL shows reliable convergence of core evidence and recovers clinically plausible dominant factors and cohort-relevant interaction patterns. Semantic steering broadens the exploration path while preserving stable dominant factors and redirecting evidence acquisition toward alternative interaction neighborhoods, supporting evidence-grounded discovery that yields auditable cohort-level insights. Code is available at https://github.com/liangyuxuan1/EGL.

Keywords: Clinical knowledge discovery · Evidence-grounded graph learning · LLM-steered semantic control · CVD outcome modeling

1 Introduction

Accurate assessment and stratification of cardiovascular disease (CVD) risk and outcomes is central to clinical decision making and preventive medicine. Modern machine learning models integrate heterogeneous clinical, imaging, and demographic variables and can achieve strong predictive performance [13,2,23]. Yet in practice, a clinical model is also expected to yield reusable clinical knowledge: which factors are reproducible, what interactions manifest stably, and when canonical associations break down across contexts. This perspective aligns with evidence-based medicine [18] and motivates clinically adopted, evidence-derived profiles such as the Framingham general CVD risk score [5].

⋆ Corresponding author.

This goal is poorly served by purely performance-driven, high-dimensional risk prediction. Modern learners can exploit subtle correlations and cohort-specific proxies that do not reliably carry over across data splits, sites, or clinically important subpopulations. Such fragility can lead to clinically meaningful failures, including misleading conclusions from standard evaluation protocols when subgroup sizes are small and uncertainty is high [11,14,7,17].

We propose Evidence-Grounded Learning (EGL), a paradigm for clinical knowledge learning that targets an explicit evidence ledger rather than a single selected model. EGL learns a distribution over feature-subset hypotheses under an estimation-of-distribution framework, maintaining a graph-structured distribution over binary inclusion vectors to capture both marginal inclusion tendencies and pairwise co-inclusion structure. Concretely, EGL parameterizes an Ising-form policy with node biases and pairwise couplings over a binary mask, and samples candidate subsets by sequential Gibbs updates: each feature bit is resampled from its conditional Bernoulli probability given the current neighborhood and the pairwise terms, yielding correlated inclusion patterns that reflect putative interactions [19,8,1]. After evaluating a sampled subset, EGL updates the Ising parameters with a REINFORCE score-function estimator [20] to increase the likelihood of high-reward subsets. In this formulation, stable clinical knowledge corresponds to node and edge structures that repeatedly receive paired evidence support through stochastic sampling and evaluation.

In realistic CVD cohorts, EGL alone can still be unstable: as a score-driven optimizer, it may converge to different local optima, and because its reward aggregates quantitative objectives (performance, confirmation pressure, coverage repair) without feature semantics, it can allocate a fixed probe budget along numerically favorable yet clinically uninformative trajectories. Prior work injects inductive biases through hand-crafted priors or task-specific modules [21,24,25,3] or uses vision–language pretraining to enable prediction and question answering [13,12,10], but the resulting knowledge is often implicit and not consolidated as an auditable, reusable evidence representation. Recent efforts couple large language models (LLMs) with optimization or search to guide exploration [22,16,9]; here, we employ LLMs as an auditable semantic control plane that steers subset sampling and probe pairing toward clinically meaningful exploration, while leaving estimation and confirmation to the underlying evidence machinery.

This study proposes a novel framework for diagnostic knowledge discovery that integrates (i) an Ising-style distribution over feature subsets to represent both marginal and pairwise structure, (ii) an online evidence graph (ledger) that accumulates reproducible node/edge statistics from repeated cross-validated evaluations, and (iii) an LLM-based plane that provides semantic steering under uncertainty and budget constraints. We evaluate the proposed framework on COPDGene [15] to perform evidence-grounded CVD knowledge discovery from CT-derived biomarkers and clinical variables. Empirically, it recovers clinically oriented and more reproducible features and interactions than purely numerical search, and yields an evidence-grounded graph representation that supports interpretable routing rules, which are promising for medical applications.

2 Method

2.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a cohort, where $x_i \in \mathbb{R}^d$ collects heterogeneous candidates (clinical variables, CT-derived biomarkers, demographics) and y_i is the CVD outcome. Standard pipelines typically optimize a point estimate (a single feature set and a single optimal model), which can be fragile under correlated covariates, sampling variability, and model stochasticity. In clinical prediction such as risk modeling, this instability is widely recognized as a key barrier to interpretability and reuse, motivating practices that emphasize robustness checks, resampling-based assessment, and transparent reporting beyond average discrimination [6,4].

Goal. We target diagnostic knowledge discovery in a sense of a distribution, beyond one final subset of selected features. Concretely, an explicit, auditable graph of reusable statements is sought: (i) nodes (features) supported by reproducible marginal evidence, and (ii) edges (feature pairs) supported by reproducible interaction evidence. The knowledge is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over candidate variables, and evidence is quantified via repeated evaluations under controlled probes.

EGL (evidence-grounded discovery). EGL learns a distribution over binary inclusion vectors $z \in \{0,1\}^{|\mathcal{V}|}$ to encode marginal inclusions (nodes) and pairwise co-inclusions (edges). Across iterations, EGL repeatedly samples candidate subsets, evaluates them, and accumulates node/edge evidence into an online ledger (the evidence graph). Stable knowledge corresponds to structures whose evidence persists and strengthens under repeated, comparable evaluations.

EGL–LLM (semantic control plane). EGL alone is score-driven and cannot express why specific nodes/edges should be tested next. We therefore introduce EGL–LLM, which keeps EGL's evidence estimation unchanged but adds an LLM control plane that outputs low-dimensional, interpretable control signals to steer the discovery trajectory toward clinically meaningful exploration under a finite evaluation budget.

Fig. 1 summarizes the EGL loop and how the LLM control plane injects a semantic agenda to steer (i) subset sampling and (ii) paired probing, while keeping the evidence estimands unchanged.

2.2 EGL: Two-layer Knowledge Graph

Sampling policy (Ising-form subset distribution). A distribution $\pi_\psi(z)$ over inclusion vectors $z \in \{0,1\}^{|\mathcal{V}|}$ is used to sample candidate subsets, where $\psi = \{b_0, \theta, \phi\}$: b_0 is a global bias controlling typical subset size, θ_f parameterizes the marginal inclusion tendency of feature f, and ϕ_{uv} parameterizes pairwise co-inclusion structure for $(u, v) \in \mathcal{E}$. Candidate subsets are drawn by Gibbs updates under this Ising-form policy [19,8,1], where binary inclusion variables are resampled conditioned on the current mask and pairwise co-inclusion terms. At each iteration, multiple subsets $S \subseteq \mathcal{V}$ are drawn and evaluated under the same

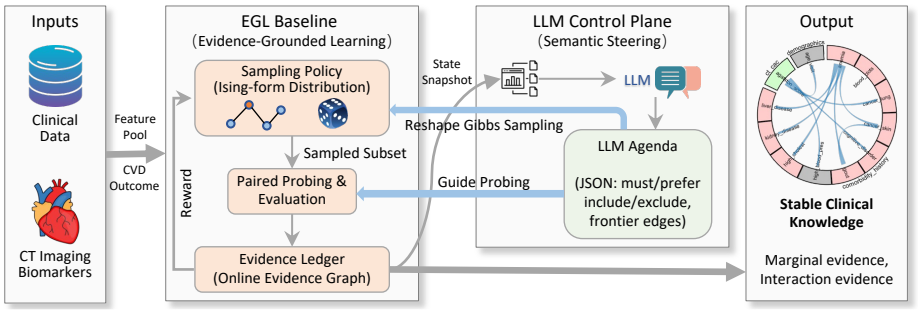

Fig. 1. Workflow for LLM-steered clinical knowledge discovery. Clinical variables and CT-derived imaging biomarkers define a feature pool for CVD outcome modeling. EGL iteratively samples feature subsets from an Ising-form policy, evaluates them with paired probes, and accumulates node/edge evidence in an online ledger, which in turn drives score-function updates of the sampling policy. EGL–LLM augments EGL with an LLM control plane: a compact state snapshot is mapped to a strict JSON agenda that is applied to guide paired probing and reshape Gibbs sampling, producing stable clinical knowledge as reproducible node (marginal) and edge (interaction) evidence.

probe budgets, and ψ is updated from score-function (REINFORCE-style) policy gradients computed from their per-subset rewards [20]. This update treats each reward as the stochastic return of a sampled mask, while accumulated support and uncertainty signals temper how strongly candidate interactions influence subsequent sampling.

Paired probing, evaluation, and reward. Given an inclusion state z sampled from the policy (equivalently, a feature subset $S = \{f \in \mathcal{V} : z_f = 1\}$), a small set of node and edge probes is selected within S. For each probed node $f \in S$, matched counterfactual evaluations are constructed by comparing S to $S \setminus \{f\}$. For each probed edge $(u, v) \subseteq S$, a matched 2×2 block is constructed by evaluating the four subsets induced by removing none, one, or both endpoints from S. Each subset is evaluated by a downstream predictor (logistic regression in this study) with K-fold cross-validation to obtain a metrics summary $\mathbf{m}(\cdot)$ (e.g., AUC, AuPRC, sensitivity, specificity, precision, recall). The probe outcomes are then converted into a scalar reward $\mathcal{R}(S)$ for score-function policy learning: a confirmation term favors probes with high confidence (absolute t-statistics) weighted by their current trust (support-derived), and an acquisition term rewards positive trust gain induced by the newly added paired probes. Node probes quantify individual-variable gain, whereas edge probes quantify joint gain beyond separate contributions. In addition, a barrier penalty based on calibrated instability and operating-point floors rejects clinically unacceptable subsets regardless of probe outcomes.

Evidence ledger (online evidence graph). All paired probe outcomes are accumulated into an online ledger, yielding auditable node and edge estimands.

For a node $f \in \mathcal{V}$, the ledger records the marginal effect

$$\Delta_f \;=\; \mathbb{E}[m(\cdot) \mid z_f{=}1] - \mathbb{E}[m(\cdot) \mid z_f{=}0]\,, \tag{1}$$

and for an edge $(u, v) \in \mathcal{E}$, the ledger records an interaction effect via difference-in-differences,

$$\begin{aligned} \Gamma_{uv} = \Big(&\mathbb{E}[m(\cdot) \mid z_u{=}1, z_v{=}1] - \mathbb{E}[m(\cdot) \mid z_u{=}1, z_v{=}0]\Big) \\ &-\Big(\mathbb{E}[m(\cdot) \mid z_u{=}0, z_v{=}1] - \mathbb{E}[m(\cdot) \mid z_u{=}0, z_v{=}0]\Big), \end{aligned} \tag{2}$$

where $m(\cdot)$ is defined as a weighted combination of normalized AUC and AuPRC. Repeated paired probes provide uncertainty proxies (standard errors and absolute t-statistics) and support counts (paired exposures) for both Δ_f and Γ_{uv}.

2.3 LLM Control Plane: Semantic Steering

Purely numerical distribution learning may over-test locally strong but clinically redundant regions. We therefore use the LLM as a semantic scheduler: once per generation, it reads a compact discovery summary and outputs an auditable agenda that biases which subsets and paired probes are tested next, without proposing the final feature set or modifying the ledger-defined estimands.

Concretely, the controller is prompted with (i) a structural snapshot of the current evidence ledger highlighting supported vs. evidence-starved nodes/edges (support counts and uncertainty proxies), (ii) phase and progress diagnostics from recent iterations (e.g., stability trends, variability, and failure-mode frequencies), (iii) the current sampled set to ensure feasibility of proposed probes, and (iv) a feature glossary that provides canonical names, clinical domains, brief meanings, and missingness cues for semantic plausibility and feasibility. The controller returns a strict JSON agenda with a small number of fields (`must_include`, `prefer_include`, `prefer_exclude`, `frontier_edges`), where the inclusion/exclusion lists provide soft preferences, and frontier edges nominate under-supported, semantically plausible interaction candidates. The agenda is then applied through two named control paths: Agenda-Shaped Gibbs Sampling (AS-Gibbs), which biases subset sampling by reshaping Gibbs conditionals via odds-calibrated logit biases, and Agenda-Guided Probe Pairing (AG-Probe), which prioritizes which paired probes are instantiated for evidence acquisition. Both paths are enabled in the reported experiments.

3 Experiments

Dataset. We used the COPDGene cohort [15], a smokers/chronic obstructive pulmonary disease (COPD)-enriched multi-site cohort with CT scans and clinical variables, as our primary dataset, because it provides genomic measurements that enable future extension of the proposed discovery pipeline to gene-informed risk modeling. We restricted the analysis to participants with available chest CT

scans. CT-derived cardiac phenotypes were extracted using the cardiac segmentation and quantification pipeline developed by Xu *et al.* [21], yielding features that summarize cardiac chambers, great vessels, myocardial wall, peri-cardiac fat, and coronary artery calcium (CAC). In parallel, we curated structured variables from COPDGene clinical information, covering demographics, vitals, smoking-related factors, comorbidity history, and medications. To improve robustness, CT features were screened via outlier handling using the 1st/99th percentile range, and clinical variables were filtered by missingness with a 70% missing-rate threshold. The final feature set contains 75 variables in total (21 CT-derived and 54 clinical).

CVD outcome labels were defined using COPDGene-reported cardiovascular events spanning coronary artery disease, coronary revascularization, peripheral vascular disease, myocardial infarction/heart attack, cerebrovascular events, and congestive heart failure. A participant was labeled as CVD-positive if any of these events/diagnoses was present, following criteria reviewed with clinicians. To align with a status/outcome modeling setting, we constructed one sample per participant by selecting a single visit record under the outcome definition, and performed downstream modeling on this subject-level cohort. The resulting cohort includes 9,638 participants (2,090 positives; 7,548 negatives), with prevalence 0.2168.

Experimental setup. EGL and EGL–LLM were executed for the same number of iterations under an identical evaluation budget per iteration, using 10 paired node probes and 30 paired edge probes. This ensures that any differences in convergence or discovered structures are not attributable to unequal evidence acquisition cost. For EGL–LLM, the controller was a locally deployed Qwen3–14B–Instruct model. To control stochasticity, the discovery code used the same fixed random seed across methods. Consequently, the primary additional source of randomness in EGL–LLM arises from the LLM-generated agenda, so observed trajectory differences are attributable to semantic control rather than uncontrolled sampling noise.

Stability and search trajectory. Fig. 2(a–b) report evidence convergence to the final snapshot by tracking, at each iteration, the Spearman correlation between the current and final core-evidence rankings (core nodes/edges selected by confidence $|t|$). This diagnostic quantifies how quickly the ordering of the most supported effects becomes consistent with the final discovery outcome. Two trends are observed. First, edge evidence converges substantially more slowly than node evidence in both methods, consistent with the combinatorial size of the interaction space relative to the node space. Second, EGL rapidly stabilizes the main node ranking, whereas EGL–LLM continues small substitutions around the same core, indicating broader exploration around a stable main-effect anchor.

Fig. 2(c) visualizes the explored subset space using a 2D Multidimensional Scaling (MDS) embedding of sampled candidate subsets, overlaid with density contours. The two methods share a high-density core region, but EGL–LLM shows a lower peak density and broader coverage across the embedding, suggesting a more dispersed exploration trajectory. This broader footprint is consistent

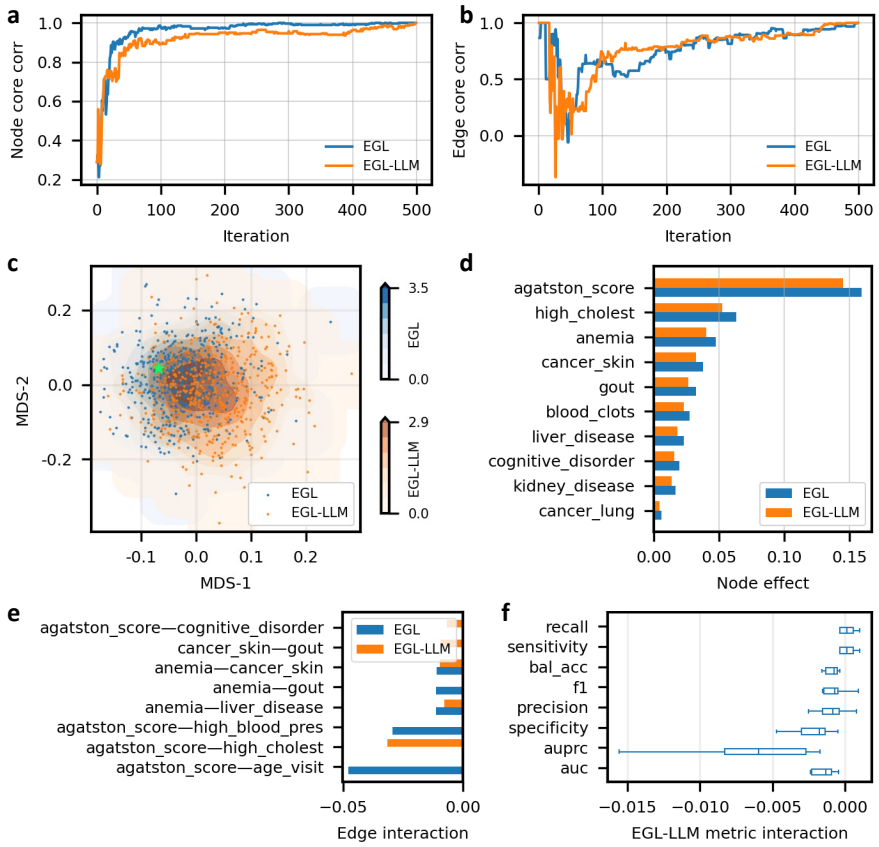

Fig. 2. EGL vs. EGL–LLM on COPDGene-based CVD outcome knowledge discovery. (a,b) Convergence diagnostics: Spearman correlation between the $|t|$-based core-evidence ranking at each iteration and the final snapshot, shown for nodes (a) and edges (b) using the top 15 by confidence $|t|$. (c) Search footprint: 2D MDS embedding of all sampled subsets across iterations with density contours; the green star marks the shared initialization. (d) Final top-10 effective nodes by confidence $|t|$, visualized by estimated node effects Δ_f. (e) Final top-5 edge interactions by confidence $|t|$, visualized by interaction effects Γ_{uv}. (f) Metric-level interaction deltas for the top-15 edges of EGL–LLM.

with the local substitutions in Fig. 2(a), while Fig. 2(b) shows that edge-evidence consolidation is not reduced.

Top effective nodes and interactions. Fig. 2(d–e) summarizes the most evidence-supported node effects (Δ_f) and edge interactions (Γ_{uv}) at the final snapshot. Both methods recover a shared core of clinically plausible factors. In Fig. 2(d), the strongest node effect is the Agatston score (CAC), followed by established cardiovascular risk correlates such as hypercholesterolemia, and thrombotic history (`blood_clots`). The remaining top nodes are dominated by

comorbidity indicators (`anemia`, `gout`, `kidney_disease`, `liver_disease`, and selected cancer histories), consistent with COPDGene being an older, multimorbid cohort where systemic disease burden and inflammatory/metabolic dysregulation can track cardiovascular status beyond canonical risk factors. Importantly, the top-10 node set is identical for EGL and EGL–LLM, indicating that the core main-effect evidence is robust to the search controller under a fixed probe budget. In Fig. 2(e), the strongest supported interactions are predominantly negative, with several involving Agatston score (e.g., Agatston score with high cholesterol, with high blood pressure, and with age at visit). It suggests a diminishing incremental benefit when combining CAC with correlated clinical markers in this cohort and model, where CAC already captures substantial downstream risk-relevant variation.

Differences and their implications. While the dominant node effects coincide, the highest-confidence interaction set differs between EGL and EGL–LLM. This is the main empirical effect of semantic steering: it changes which interaction neighborhoods receive repeated evaluation while preserving the dominant node evidence. This contrast is expected because interaction evidence is harder to consolidate than node evidence: the candidate edge space is much larger, and Γ_{uv} is a second-order estimand that is more sensitive to which paired probes are allocated under limited evaluations. EGL–LLM therefore shapes the route by which interaction evidence is accumulated, rather than overturning the core main-effect conclusions.

Edge interactions emphasize redundancy. Fig. 2(f) summarizes metric-level interaction deltas for the top-15 EGL–LLM edges. The interactions are predominantly negative, indicating sub-additive contributions when combining correlated factors under the paired-probe estimand. Across metrics, the largest and most consistent degradation is observed for AuPRC, while changes in other metrics are comparatively small, suggesting that redundancy among the discovered interaction pairs primarily affects performance in the low-prevalence regime, where precision–recall behavior is most sensitive to overlapping evidence and false positives.

Limitations and future work. The current evaluation is limited to a single cohort. COPDGene is a strong primary testbed because it is large, multi-site, and combines CT-derived cardiac/CAC phenotypes with clinical risk factors, comorbidities, medications, and CVD labels; however, multi-cohort validation is needed before interpreting the discovered graph as population-level CVD knowledge. Such validation requires harmonizing imaging pipelines, clinical variables, and outcome definitions across cohorts. We also do not directly benchmark against feature selection or interaction discovery methods. Because these methods return selected variables, explicit interaction terms, model explanations, or dependency structures rather than repeated node/edge evidence, a fair comparison requires a protocol that maps their outputs into comparable node- and edge-level evidence.

4 Conclusion

This work proposed EGL as a paradigm for clinical knowledge discovery, framing discovery as learning an evidence-grounded graph rather than selecting a single best feature set. EGL combines an online paired-evidence ledger for node and edge effects with an Ising-form subset distribution updated by score-function learning. EGL–LLM further adds an auditable LLM control plane that steers subset sampling and probe pairing under the same estimands and probe budget. On COPDGene CVD outcome modeling with CT-derived cardiac phenotypes and clinical variables, the method shows stable convergence of core evidence, recovers clinically plausible dominant factors, and reveals predominantly negative, redundancy-oriented interaction structure among correlated risk markers. Semantic steering redirects evidence acquisition toward alternative interaction neighborhoods while preserving the dominant node core. The current study remains cohort-level; multi-cohort validation and fair comparisons with feature-selection or interaction-discovery methods require additional harmonization protocols and are important future work.

Acknowledgments. This work was partly supported by the National Science Foundation through the Faculty Early Career Development Program (CAREER) under Award 2046708.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann machines. Cognitive Science **9**(1), 147–169 (1985). https://doi.org/10.1016/S0364-0213(85)80012-4
2. Chao, H., Shan, H., Homayounieh, F., Singh, R., Khera, R.D., Guo, H., Su, T., Wang, G., Kalra, M.K., Yan, P.: Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. Nature Communications **12**(1), 2963 (2021). https://doi.org/10.1038/s41467-021-23235-4
3. Chen, A., Li, Y., Qian, W., Morse, K., Miao, C., Huai, M.: Modeling and understanding uncertainty in medical image classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15010, pp. 557–567. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72117-5_52
4. Chowdhury, M.Z.I., Turin, T.C.: Variable selection strategies and its importance in clinical prediction modelling. Family Medicine and Community Health **8**(1), e000262 (2020). https://doi.org/10.1136/fmch-2019-000262
5. D'Agostino Sr., R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., Kannel, W.B.: General cardiovascular risk profile for use in primary care: The framingham heart study. Circulation **117**(6), 743–753 (2008). https://doi.org/10.1161/CIRCULATIONAHA.107.699579

6. Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., Salanti, G.: Developing clinical prediction models: a step-by-step guide. BMJ **386**, e078276 (2024). https://doi.org/10.1136/bmj-2023-078276
7. Futoma, J., Simons, M., Doshi-Velez, F., Kamaleswaran, R.: Generalization in clinical prediction models: The blessing and curse of measurement indicator variables. Critical Care Explorations **3**(7), e0453 (2021). https://doi.org/10.1097/CCE.0000000000000453
8. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-6**(6), 721–741 (1984). https://doi.org/10.1109/TPAMI.1984.4767596
9. Huang, B., Wu, X., Zhou, Y., Wu, J., Feng, L., Cheng, R., Tan, K.C.: Evaluation of large language models as solution generators in complex optimization. IEEE Computational Intelligence Magazine **20**(4), 56–70 (2025). https://doi.org/10.1109/MCI.2025.3580520
10. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems. vol. 36, pp. 28541–28564. Curran Associates, Inc. (2023), https://papers.nips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html
11. Liang, Y., Chao, H., Zhang, J., Wang, G., Yan, P.: Unbiasing fairness evaluation of radiology AI model. Meta-Radiology **2**(3), 100084 (2024). https://doi.org/10.1016/j.metrad.2024.100084
12. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., Soliman, A., Chen, C., Ding, T., Wang, J.J., Gerber, G., Liang, I., Le, L.P., Parwani, A.V., Weishaupt, L.L., Mahmood, F.: A multimodal generative AI copilot for human pathology. Nature **634**(8033), 466–473 (2024). https://doi.org/10.1038/s41586-024-07618-3
13. Niu, C., Lyu, Q., Carothers, C.D., Kaviani, P., Tan, J., Yan, P., Kalra, M.K., Whitlow, C.T., Wang, G.: Medical multimodal multitask foundation model for lung cancer screening. Nature Communications **16**(1), 1523 (2025). https://doi.org/10.1038/s41467-025-56822-w
14. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 151–159. CHIL '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3368555.3384468
15. Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., Crapo, J.D.: Genetic epidemiology of COPD (COPDGene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease **7**(1), 32–43 (2010). https://doi.org/10.3109/15412550903499522
16. Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M.P., Dupont, E., Ruiz, F.J.R., Ellenberg, J.S., Wang, P., Fawzi, O., Kohli, P., Fawzi, A.: Mathematical discoveries from program search with large language models. Nature **625**(7995), 468–475 (2024). https://doi.org/10.1038/s41586-023-06924-6
17. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

18. Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn't. BMJ **312**(7023), 71–72 (1996). https://doi.org/10.1136/bmj.312.7023.71
19. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning **1**(1–2), 1–305 (2008). https://doi.org/10.1561/2200000001
20. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning **8**(3–4), 229–256 (1992). https://doi.org/10.1007/BF00992696
21. Xu, X., Budoff, M.J., Kalra, M.K., Wang, G., Yan, P.: CAC-MAE: A calcification-aware masked autoencoder for cardiovascular disease risk assessment on low-dose CT. In: Machine Learning in Medical Imaging – MLMI 2025. Lecture Notes in Computer Science, vol. 16241, pp. 308–317. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-09513-8_30
22. Yang, T., Yang, T., Lyu, F., Liu, S., Liu, X.: ICE-SEARCH: A language model-driven feature selection approach. arXiv preprint arXiv:2402.18609 (2024), https://arxiv.org/abs/2402.18609
23. Yang, Z., Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Cardiovascular disease detection from multi-view chest x-rays with BI-Mamba. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15005, pp. 134–144. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72086-4_13
24. Yu, L., Min, W., Wang, S.: Boundary-aware gradient operator network for medical image segmentation. IEEE Journal of Biomedical and Health Informatics **28**(8), 4711–4723 (2024). https://doi.org/10.1109/JBHI.2024.3404273
25. Zheng, Z., Hayashi, Y., Oda, M., Kitasaka, T., Mori, K.: A bayesian approach to weakly-supervised laparoscopic image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15006, pp. 14–24. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72089-5_2