Pathologic Complete Response in Lung Cancer from Structured Clinical Records: A Retrospective Interpretability Study

Ciro Russo[1,2][0009−0002−8751−8605], Giulio Russo[1,2][0009−0006−5083−7548], Domenico Paolo[1,2][0009−0001−8997−2839], Valerio Guarrasi[3][0000−0002−1860−7447], Alice Natalina Caragliano[3][0009−0007−8097−7290], Carlo Greco[4][0000−0002−1068−5045], Alessio Cortellini[5][0000−0002−1209−5735], Marco Russano[5][0000−0002−6963−3864], Sara Ramella[4][0000−0002−5782−7717], Alessandro Bria[1,2][0000−0002−2895−6544], Paolo Soda[6][0000−0003−2621−072X], and Claudio Marrocco[1,2][0000−0003−0840−7350]

[1] Department of Electrical and Information Engineering, University of Cassino and Lazio Meridionale, 03043 Cassino (FR), Italy
[2] EUT+, European University of Technology, European Union, Italy
[3] Unit of Artificial Intelligence and Computer Systems, Department of Engineering, Universita' Campus Bio-Medico di Roma, Rome, Italy
[4] Research Unit of Radiation Oncology, Dept. of Medicine and Surgery, Universita' Campus Bio-Medico di Roma, Italy
[5] Operative Research Unit of Medical Oncology, Fondazione Policlinico Universitario Campus Bio-Medico, Italy
[6] Department of Diagnostics and Intervention, Umeå University, Sweden

Abstract. Pathologic complete response (pCR) is a recognized surrogate of long-term survival in non-small-cell lung cancer (NSCLC). Through a retrospective analysis, this study investigates which structured features of the Electronic Health Record (EHR) are associated with pCR, how stable these associations are across different inductive biases, and whether the patient cohort exhibits identifiable subgroup structure. On a cohort of 101 NSCLC records, we compare Logistic Regression, Random Forest and XGBoost under a stratified group cross-validation protocol grouped by patient identifier. Each best estimator is subjected to a comprehensive interpretability analysis combining absolute Pearson correlation, mutual information, univariate tests and permutation importance into a consensus ranking, complemented by SHAP, Recursive Feature Elimination, k-Means subgroup discovery and a pairwise interaction screen. Logistic Regression attains the best test MCC (0.372). The consensus ranking consistently identifies a recurring set of features related to lymph-node clearance, post-surgical staging, treatment regimen and histology across all three models, indicating stable associations regardless of the inductive bias. The subgroup analysis links the lowest-pCR cluster to incomplete molecular profiling. These findings indicate that structured EHR data, analyzed through interpretable models, yields consistent associations with pCR in lung cancer.

Keywords: Pathologic complete response · Non-small-cell lung cancer · Electronic Health Records · Explainable machine learning · Feature selection

1 Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide, accounting for more than 1.38 million deaths every year [9]. Non-small cell lung cancer (NSCLC) represents $80-85\%$ of all cases and its prognosis depends strongly on stage at diagnosis, with five-year overall survival ranging from 92% for small, localized lesions to 26% for invasive disease [7]. Neoadjuvant chemotherapy, chemo-radiation [14] and immune-checkpoint inhibitors [10] are the cornerstones of treatment, and their efficacy is commonly summarized by the *pathologic complete response* (pCR), i.e. the absence of residual malignant cells in the resected specimen. pCR is a recognized surrogate of long-term survival, and the identification of clinical features associated with pCR may support therapy planning and the stratification of patients with different response profiles.

Electronic Health Records (EHRs) collect structured items (demographics, staging, molecular markers, treatment regimens, surgical procedures, toxicity events) together with unstructured free-text narratives [18]. Prior work combining both sources achieved an MCC of 0.371 on a cohort of 67 NSCLC patients [15]; notably, structured variables alone already reached an MCC of 0.333, suggesting that a substantial fraction of the association signal resides in the structured component. Predictive performance, however, is only one dimension along which a clinical decision-support tool should be evaluated. In precision oncology, cohorts are small and heterogeneous, and clinicians need to understand *which* variables drive a prediction and *whether* the learned reasoning is consistent with established medical knowledge [1]. A single global score is unlikely to reach the bedside without transparent, clinically inspectable evidence. Tabular classifiers such as Decision Trees, Random Forests, SVMs and gradient-boosted ensembles are the historical workhorse of oncological prognostic modeling on structured EHR data and remain competitive with deeper architectures on small clinical cohorts [6,17]. pCR has been studied as a surrogate endpoint across multiple cancer types [5,4,16], yet most approaches concentrate on raw predictive performance and devote limited attention to interpretability. Univariate statistical tests, mutual information [11], permutation importance [2], SHAP [13] and Recursive Feature Elimination [8] are individually applied in various clinical prediction tasks, but a systematic analysis consolidating these methods into a consensus ranking, complemented by subgroup discovery and pairwise interaction screening, has not been applied to pCR prediction in NSCLC from structured EHR data.

Motivated by this gap, the present study focuses exclusively on structured EHR features and revisits pCR prediction on an updated institutional cohort of 101 NSCLC records. We compare Logistic Regression, Random Forest and XGBoost under a stratified group cross-validation protocol grouped by patient

identifier, and apply a suite of complementary interpretability techniques consolidated into a consensus ranking. We further cluster patients into subgroups with potentially different response patterns and screen pairwise feature interactions to surface candidate combination biomarkers. Together, these retrospective analyses investigate which structured EHR features are consistently associated with pCR across models and whether the cohort exhibits identifiable subgroup structure

2 Materials and Methods

We analyze the structured clinical features extracted from the EHR of an institutional cohort, training three classifiers under a patient-level cross-validation protocol and probing them with a systematic interpretability analysis.

2.1 Dataset and Protocol

The dataset is drawn from the EHR system of a single oncological center and comprises $N = 101$ records belonging to NSCLC patients treated with neoadjuvant therapy followed by surgical resection, with a positive pCR rate of 34.65%. This cohort is an updated and extended release of a previously described collection [15], which originally included 67 patients. The analysis is retrospective: the feature set includes both pre- and post-surgical variables, as the objective is to identify stable associations with pCR rather than to develop a preoperative prediction tool.

For each record, $D = 192$ structured features are extracted, organized into the following clinical categories: demographics (sex, age at diagnosis) and anthropometrics (height, weight), smoking history, histological subtype (adenocarcinoma, squamous cell carcinoma, neuroendocrine, poorly differentiated), molecular marker testing status and results (EGFR, ALK, ROS1, KRAS, BRAF, RET, NTRK, HER2, MET, PDL1), clinical and pathological TNM staging, neoadjuvant treatment scheme and regimen, radiotherapy parameters, surgical procedure, lymph-node clearance, toxicity events and follow-up status. Categorical attributes are one-hot encoded; missing entries are retained as a dedicated level rather than imputed or discarded, so that incomplete molecular profiling is itself informative.

A 21-record held-out test set (20% of the records) is drawn uniformly at random with a fixed seed and set aside before any model fitting. Hyperparameters are tuned on the remaining 80 records via 5-fold stratified group cross-validation, which simultaneously groups records by patient identifier and stratifies by target so that no patient appears in more than one fold. The selected estimator is then refitted on the full training set and evaluated on the untouched test fold; given the limited test set size, all reported metrics should be interpreted as indicative and consistent with the exploratory nature of this study.

2.2 Classifiers and Metrics

Three classifiers with distinct inductive biases are compared: Logistic Regression (LR), a linear baseline with calibrated probabilities; Random Forest (RF), a bagged ensemble; and XGBoost (XGB), a gradient-boosted ensemble. Each model is tuned via exhaustive grid search over a predefined hyperparameter space: LR over the regularization strength $C \in \{0.01, 0.1, 1, 10, 100\}$ with L2 penalty and balanced class weights; RF over the number of trees $\in \{100, 200, 500\}$, maximum depth $\in \{\text{None}, 10, 20, 30\}$, minimum samples per split $\in \{2, 5, 10\}$ and minimum samples per leaf $\in \{1, 2, 4\}$; XGB over the number of boosting rounds $\in \{100, 200, 400\}$, maximum depth $\in \{3, 6, 10\}$, learning rate $\in \{0.01, 0.1, 0.3\}$ and subsample ratio $\in \{0.6, 0.8, 1.0\}$. The optimization target is the Matthews Correlation Coefficient (MCC), which remains balanced under class imbalance [3], and evaluated on both the training-validation phase (CV MCC) and the independent test set. On the test set we additionally report Accuracy, Precision, Recall, F1, ROC-AUC and the confusion matrix at the default 0.5 threshold.

2.3 Interpretability Analysis

Once the best estimator $\hat{f}$ is fixed for each classifier, we apply a suite of eight complementary analyses organized into three groups: model-agnostic statistical methods, model-based attributions, and structural analyses. Four of these methods are consolidated into a consensus ranking.

Model-agnostic statistical methods Three univariate analyses capture different forms of feature–target association independently of the fitted model. *(i) Absolute Pearson correlation* measures linear associations between each feature and the binary target. *(ii) Mutual information*, estimated through the Kraskov–Stögbauer–Grassberger algorithm [11], detects non-linear and threshold effects that linear measures miss. *(iii) Univariate hypothesis tests* provide a per-feature score and p-value: the chi-square test is used when all features in the training matrix are non-negative (as is the case after one-hot encoding), and the ANOVA F-test otherwise.

Model-based attributions Three techniques exploit the structure of the fitted model. *(iv) Permutation importance* [2] repeatedly shuffles each feature on the test set and measures the induced drop in ROC-AUC, producing a global, model-specific ranking. *(v) SHAP* [13] distributes each prediction across features through model-aware explainers, producing both a global ranking and per-instance attributions; SHAP outputs are reported alongside the consensus but not aggregated into it. *(vi) Recursive Feature Elimination (RFE)* [8] iteratively prunes the least important feature until a parsimonious panel of $K = 20$ features remains.

Structural analyses Two methods look beyond single-feature scores. *(vii) Subgroup discovery:* training records are clustered with k-Means ($k = 3$, chosen as a minimal partition that allows comparison across response levels while remaining interpretable on a small cohort). This method is retained as an exploratory tool despite the predominantly binary feature space, as the goal is to identify broad response patterns rather than geometrically precise clusters. For tree-based estimators, dominant per-cluster features are obtained by refitting a clone of $\hat{f}$ within each cluster. *(viii) Pairwise interaction screen:* for the top $K = 15$ features ranked by the model, the element-wise product $\mathbf{X}_{:,j_1} \odot \mathbf{X}_{:,j_2}$ is substituted for column j_1 and the absolute change in score is measured for each pair (j_1, j_2); the training set is used to avoid the high variance that would result from evaluating interactions on the 21-record test fold. Pairs exceeding a threshold $\omega > \tau = 0.01$ are retained as candidate interactions. For LR, which does not provide a native feature ranking, the first $K = 15$ columns of the encoded matrix are used instead.

Consensus ranking The four per-feature scores from methods (i)–(iv) live on different scales. We convert each into a feature-wise rank $r_{m,j}$ and define the consensus as $\bar{r}_j = \frac{1}{M} \sum_{m=1}^{M} r_{m,j}$ with $M = 4$, then re-rank $\bar{r}_j$ to obtain the final position of each feature. This rank-based aggregation is robust to outliers from any single method and stabilizes the identification of reliable features in the low-sample regime [12]; the RFE selection mask is reported as a supplementary indicator. Note that Pearson correlation and univariate tests are not fully independent on binary feature spaces, which may reduce the effective diversity of the consensus.

3 Experiments and Results

We evaluate the three classifiers on the institutional cohort under the protocol described in Sec. 2 and organize the results around four research questions: predictive performance (RQ1), stability of feature attributions across models (RQ2), cohort heterogeneity (RQ3) and pairwise feature interactions (RQ4).

3.1 Predictive Performance

Table 1 reports the performance of the three classifiers on the institutional cohort. LR achieves the highest test MCC (0.372) and F1 (0.571), with a linear model whose coefficients are directly interpretable. RF presents a contrasting profile: it attains the highest CV MCC (0.389) and test ROC-AUC (0.740), indicating strong threshold-free discrimination, but its hard predictions at the default 0.5 threshold collapse to a near-random MCC (-0.040) and a recall of 0.125, a discrepancy consistent with probability miscalibration that would require a calibration step before clinical deployment. XGBoost achieves a moderate test MCC (0.240) and the lowest test ROC-AUC (0.625).

Table 1. Performance of the three classifiers on the institutional cohort. Best value per column is in bold.

Model	CV MCC	Test MCC	ROC-AUC	Acc.	F1	Recall
LR	0.245	**0.372**	0.664	0.714	0.571	0.500
RF	0.389	−0.040	0.740	0.571	0.182	0.125
XGB	0.369	0.240	0.625	0.667	0.364	0.250

Table 2. Top features by consensus ranking (lower = more informative). Features in the top-20 of all three runs are marked with ⋆.

Feature	LR	RF	XGB
PN_CLEARANCE_LINFONODALE_SI ⋆	1.75	1.75	1.75
PTNM_YPT0N0 ⋆	10.00	4.00	45.25
RET_non eseguito ⋆	39.38	9.50	10.88
DIAGNOSI_Carcinoma Squamoso	16.50	—	56.75
BOOST_SI ⋆	19.12	24.12	35.38
RISPOSTA_SD ⋆	28.50	36.00	49.75
HER2_non eseguito	43.50	40.50	—
EGFR_negativo	23.50	23.00	—
KRAS_negativo	—	21.88	42.12
PTNM_PT3N0 ⋆	52.00	25.38	41.38

3.2 Stability of Feature Attributions

Table 2 reports the top entries of the consensus ranking across the three classifiers. The lymph-node-clearance indicator *PN_CLEARANCE_LINFONODALE_SI* ranks first for all three models, with a consensus value of 1.75, an absolute Pearson correlation of 0.573, a mutual-information score of 0.155, and RFE selection in every run. Other features recurring in the top positions across all three rankings include the post-surgical staging codes *PTNM_YPT0N0* and *PTNM_PT3N0*, the absence of RET testing (*RET_non eseguito*), the radiotherapy boost (*BOOST_SI*) and the stable disease label (*RISPOSTA_SD*). The high ranking of *PTNM_YPT0N0*, the post-surgical pathological staging code corresponding to pCR itself, reflects its direct association with the target rather than independent predictive information, and should be interpreted accordingly. Two further features appear in the top-20 of two out of three rankings: squamous-cell histology (*DIAGNOSI_Carcinoma Squamoso*, LR and XGB) and untested HER2 (*HER2_non eseguito*, LR and RF). Permutation importance reveals that XGB relies on a sparser subset of features than LR and RF.

3.3 Cohort Heterogeneity and Feature Interactions

k-Means with $k = 3$ partitions the training set into three subgroups of size 18, 55 and 7 patients, with empirical pCR rates of 38.89%, 32.73% and 28.57%, respectively. The ordering by pCR rate is monotonic across clusters. The smallest cluster ($\mathcal{C}_3$) limits the robustness of per-cluster claims, yet inspection of the per-cluster feature means indicates that $\mathcal{C}_3$ is enriched in advanced staging codes and

in missing molecular marker values, suggesting that the lowest-pCR subgroup is also the one with the highest degree of incomplete molecular profiling.

The pairwise interaction screen exhibits a strongly model-dependent pattern. For LR, which does not provide a native feature ranking, the analysis uses the first $K = 15$ feature columns of the encoded matrix (see Sec. 2.3), which correspond to the continuous covariates; the resulting interaction matrix is dense, with *ALTEZZA* × *SIG/DIE*, *ALTEZZA* × *ETA* and *ALTEZZA* ×*DURATA_RT* all reaching $\omega = 0.663$, and *ALTEZZA* ×*PDL1* attaining $\omega = 0.613$. These values must be interpreted with care: they indicate that perturbing any of these continuous columns shifts the LR decision surface, but they are not selected on the basis of predictive importance. XGB recovers a different picture, with interactions concentrated around squamous-cell histology: the strongest pair is *DIAGNOSI_Carcinoma Squamoso* × *TOX_ESOFAGEA_G1* ($\omega = 0.050$), followed by interactions of the same histology with chemotherapy regimen, surgical procedure and the ypT0N0 staging code [14]. RF returns a near-degenerate matrix in which all top pairs share the same first feature (*RET_non eseguito*) and have $\omega = 0$ indicating that the ensemble relies on *PN_CLEARANCE_LINFONODALE_SI* (permutation importance 0.263) plus a few additive contributors, with no detectable second-order structure.

4 Discussion and Conclusions

On small, class-imbalanced NSCLC cohorts, the primary value of a systematic analysis of structured EHR features lies in clarifying the structure of the available evidence and identifying stable associations between clinical variables and pCR. The MCC of 0.372 attained by LR matches the best result reported in prior work [15], suggesting that the current cohort size and feature set define a performance ceiling that larger cohorts, multimodal inputs (e.g. imaging or radiomics), or patient-tailored models would be required to overcome. The miscalibration of RF reinforces this interpretation: when the operating point matters — and in pCR prediction it does, since the predicted class drives treatment selection — simpler models with well-calibrated probabilities can outperform more flexible alternatives, and threshold-aware metrics such as MCC remain indispensable. The large discrepancy between CV MCC (0.389) and test MCC (-0.040) for RF further suggests overfitting on the small training fold, compounding the miscalibration effect.

The recurrence of the same top-ranked features across LR, RF and XGB is the principal empirical contribution of this work: the consensus ranking is unlikely to reflect artifacts of a specific inductive bias, and the biological plausibility of the emerging associations (lymph-node clearance, post-surgical staging, squamous histology, radiotherapy boost, response category) lends further support to their clinical relevance. The subgroup analysis is exploratory given the size of the smallest cluster (n=7), yet the observation that the lowest-pCR subgroup concentrates the highest density of missing molecular marker values constitutes a data-driven hypothesis warranting prospective investigation. The

main limitations are the cohort size (101 records, single-center); the inclusion of post-surgical features in the structured EHR, which precludes direct application in a preoperative setting and limits the scope of the analysis to retrospective association discovery; the consequent variance of the 21-record test fold; the restriction of the interaction screen to pairwise effects; and the associative rather than causal nature of all reported explanations. Finally, treatment-related features such as radiotherapy boost and response category are conditioned on the administered therapy rather than on pre-treatment patient characteristics, and their role should be interpreted with caution.

Future work will pursue four directions: cohort extension through a multicenter effort; integration of NER-driven free-text encoding alongside counterfactual explanations constrained to clinically modifiable features; introduction of probability calibration and threshold-aware training objectives; and exploration of higher-order interactions and multimodal extensions incorporating radiomics features, with the long-term objective of a prospective clinical evaluation.

Acknowledgment

This work has been supported by European union - Next Generation EU, Mission 4 Part 1 PRIN 2022 PNRR P2022P3CXJ - PICTURE (CUP D53D23021160001).

References

1. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., the Precise4Q consortium: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making **20**(1), 310 (2020). https://doi.org/10.1186/s12911-020-01332-6
2. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
3. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 6 (Jan 2020). https://doi.org/10.1186/s12864-019-6413-7
4. Crosbie, A., Loi, T., Zhang, Y., Dias, R., Awada, F., Desmedt, C., Goossens, M.: Neoadjuvant treatment and survival outcomes by pathologic complete response in HER2-negative early breast cancers. Future Oncology **19**(3), 229–244 (2023). https://doi.org/10.2217/fon-2022-0801
5. Feng, X., Song, L., Wang, S., Song, H., Chen, H., Liu, Y., Lou, C., Zhao, J., Liu, Q., Liu, Y., Zhao, R., Xing, K., Li, S., Yu, Y., Liu, Z., Yin, C., Han, B., Du, Y., Xin, R., Huang, L., Fan, Z., Zhou, F.: Accurate prediction of neoadjuvant chemotherapy pathological complete remission (pcr) for the four sub-types of breast cancer. IEEE Access (2019). https://doi.org/10.1109/ACCESS.2019.2941543
6. Forde, P.M., Spicer, J., Lu, S., Provencio, M., Mitsudomi, T., Awad, M.M., Felip, E., Broderick, S.R., Brahmer, J.R., Swanson, S.J., Kerr, K., Wang, C., Ciuleanu, T.E., Saylors, G.B., Tanaka, F., Ito, H., Chen, K.N., Liberman, M., Vokes, E.E., Taube, J.M., Dorange, C., Cai, J., Fiore, J., Jarkowski, A., Balli, D., Sausen, M., Pandya, D., Calvet, C.Y., Girard, N.: Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. New England Journal of Medicine **386**(21), 1973–1985 (2022). https://doi.org/10.1056/NEJMoa2202170

7. Goldstraw, P., Chansky, K., Crowley, J.: The iaslc lung cancer staging project: Proposals for revision of the tnm stage groupings in the forthcoming (eighth) edition of the tnm classification for lung cancer. J Thorac Oncol. (2016). https://doi.org/10.1016/j.jtho.2015.09.009
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46**(1), 389–422 (Jan 2002). https://doi.org/10.1023/A:1012487302797
9. Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., Thun, M.: Cancer statistics, 2008. CA Cancer J Clin. (2008). https://doi.org/10.3322/CA.2007.0010.
10. Jia, X.H., Xu, H., Geng, L.Y., Jiao, M., Wang, W.J., Jiang, L.L., Guo, H.: Efficacy and safety of neoadjuvant immunotherapy in resectable nonsmall cell lung cancer: A meta-analysis. Lung Cancer **147**, 143–153 (2020). https://doi.org/10.1016/j.lungcan.2020.07.001.
11. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical Review E **69**(6), 066138 (2004). https://doi.org/10.1103/PhysRevE.69.066138
12. Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv **arXiv**(arXiv:2202.01602) (Apr 2025). https://doi.org/10.48550/arXiv.2202.01602, http://arxiv.org/abs/2202.01602
13. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. arXiv **arXiv**(arXiv:1705.07874) (Nov 2017). https://doi.org/10.48550/arXiv.1705.07874, http://arxiv.org/abs/1705.07874, arXiv:1705.07874 [cs]
14. NSCLC Meta-analysis Collaborative Group: Preoperative chemotherapy for non-small-cell lung cancer: a systematic review and meta-analysis of individual participant data. Lancet **383**(9928), 1561–1571 (2014). https://doi.org/10.1016/S0140-6736(13)62159-5
15. Paolo, D., Russo, C., Russo, G., Greco, C., Cortellini, A., Russano, M., Ramella, S., Soda, P., Marrocco, C., Bria, A., Sicilia, R.: Pathologic complete response prediction with machine learning using hierarchical attention feature extraction. In: Palaiahnakote, S., Schuckers, S., Ogier, J.M., Bhattacharya, P., Pal, U., Bhattacharya, S. (eds.) Pattern Recognition. ICPR 2024 International Workshops and Challenges. p. 255–267. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-87660-8_19
16. Shu, C., Gainor, J., Awad, M., Chiuzan, C., Grigg, C., Pabani, A., Garofano, R., Stoopler, M., Cheng, S., Bhatt, N., Ravit, G.: Neoadjuvant atezolizumab and chemotherapy in patients with resectable non-small-cell lung cancer: an open-label, multicentre, single-arm, phase 2 trial. Lancet Oncology **21**(6), 786–795 (2020). https://doi.org/10.1016/S1470-2045(20)30140-6
17. Vallières, M., Braunstein, S., Ginart, J., Upadhaya, T., Woodruff, H., Zwanenburg, A., Chatterjee, A., Villanueva-Meyer, J., Valdes, G., Chen, W., Hong, J., Yom, S., Solberg, T., Löck, S., Seuntjens, J., Park, C., Lambin, P.: An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nature Cancer **2**, 709–722 (2021). https://doi.org/10.1038/s43018-021-00236-2
18. Yadav, P., Steinbach, M., Kumar, V., Simon, G.: Mining electronic health records (ehrs): A survey. ACM Comput. Surv. **50** (2018). https://doi.org/10.1145/3127881