A Time Series Foundation Model for Cancer Management

Chenlian Fu[1,2] **Ruian (Ian) Shi**[1,3] **Michele Waters**[1] **Nikolaus Schultz**[1]

Justin Jee[1] **Quaid Morris**[1] *

[1]Memorial Sloan Kettering Cancer Center, New York, NY
[2]Weill Cornell Medicine, New York, NY
[3]University of Toronto, Toronto, ON, Canada

Abstract

In cancer, predicting critical events, such as death or complications of treatment, is crucial for personalized medicine. Prediction of such events frequently ignores dense, longitudinal data such as laboratory tests and vital signs collected as part of standard of care. We present a foundation model for adverse event prediction, 'Surveillance of Patient Adverse events using Routine Clinical data' (SPARC). SPARC outperforms models based on single-timepoint measurements as well as previous time series architectures for predicting adverse cancer outcomes, i.e. side effects of chemotherapy, immunotherapy and death. SPARC generalizes to external validation datasets, excels with limited training data availability, and incorporates non-obvious features to improve outcome prediction. Overall, SPARC provides a generalizable and efficient solution for optimizing cancer treatment decisions.

1 Introduction

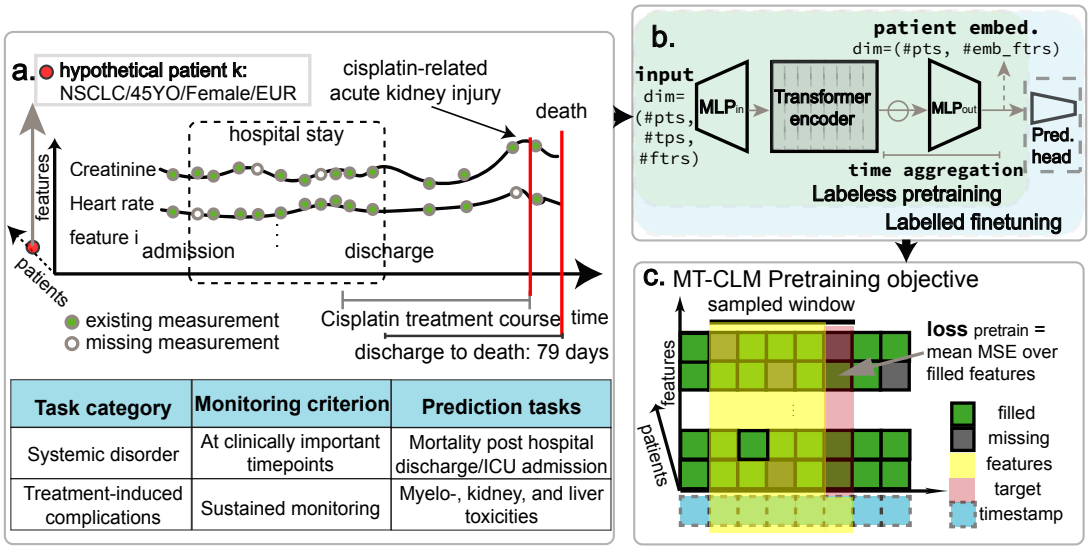
Better risk modeling can improve many aspects of cancer care. Mortality prognostication can improve utilization of hospice services and effectiveness of ICU care [1–4]. Predicting which patients might have side effects from cytotoxic chemotherapy or other antineoplastics can improve dosing and supportive medication usage for patients requiring such treatments[5, 6].

Despite progress in modeling clinical outcomes from health data, lab and vital sign measurements remain underutilized data substrates for these tasks. Prior foundation model studies have focused on billing codes, overlooking the numerical trends captured by labs and vitals [7]. Emerging works have trained foundation models on ICU time series [8–11], but ICU data differs from routine cancer care in time scale and data collection methods, limiting transferability. Moreover, lab and vital tests are central to diagnosing cancer treatment complications [12], yet existing cancer-focused models often ignore temporal aspects of such data, underuse advances in AI, or lack validation in cancer contexts.

We present a foundation model developed on institutional numerical time series data from 57,510 patients and 78,991,523 measurements, validated on extensive internal and external data, to accurately estimate risks of cancer mortality and treatment complications while identifying relevant biomarkers.

*Correspondence to {jeej, morrisq}@mskcc.org. Code available at https://github.com/tommyfuu/SPARC

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .



Task category	Monitoring criterion	Prediction tasks
Systemic disorder	At clinically important timepoints	Mortality post hospital discharge/ICU admission
Treatment-induced complications	Sustained monitoring	Myelo-, kidney, and liver toxicities

Figure 1: Overview of SPARC. Deriving clinically actionable prediction tasks from patient lab/vital time series and treatment and clinical events (a), we build SPARC with an MT-CLM pretraining objective to generate a finetunable, time-aware physiological state embedding (b,c).[2]

2 Related works

Existing time-series models for health outcome prediction have largely leveraged billing code data or continuous clinical measurement such as laboratory tests and vital signs.

Most billing code-based models have used diagnosis and procedure codes as features but omit the numerical components tied to a subset of codes, such as lab measurements [13–23], or simplify such measurements into quantiles or binary in-/out-of-range indicators [24, 25]. Despite empirical success, as reviewed and benchmarked in [13–15], these models have yet to consistently outperform supervised baselines or generalize well across hospitals.

Models that leverage continuous clinical time series data, including labs and vitals, have to-date been focused on ICU-related scenarios and tasks. Enabled by open datasets like MIMIC and PhysioNet [10, 11, 26], these models with novel architectures have achieved strong performance on ICU-specific mortality, readmission, and sepsis [27–30]. However, their utility in oncologic settings is unknown. Cancer time-series data also differs considerably from ICU data in terms of sparsity, completeness, frequency and consistency of sampling. Cancer lab tests are sampled on frequencies that vary from days to months, often contain many missing fields, and vary considerably in sampling rate through the patient's cancer journey. Recently, cancer-focused works have emerged, predicting immunotherapy response and mortality using select single-timepoint labs, and cancer-associated venous thromboembolism and early cancer risks using select lab time series, showing early promise in clinical prediction [31–34]. These trends motivate the further development of models that leverage continuous lab and vital sign time series data for predicting cancer outcomes, particularly treatment complications and mortality.

3 Methodology

3.1 SPARC model architecture and the MT-CLM pretraining objective

SPARC is a transformer encoder-based model [35] that constructs patient-specific embeddings, followed by a prediction head (e.g., binary classifier) for outcome prediction (Figure 1b). Time series data is fed in reverse order, while SPARC handles missingness via (1) padding to mask nonexistent time points and (2) a missingness token "-1" for the partially filled time points, prompting MLP_in to ignore missing data. The encoder is followed by a time aggregation module calculating min, max,

[2]#pts, tps, ftrs, ltnt_ftrs, emb_ftrs: number of patients, timepoints, features, latent-, and embedding features.

and mean over time for each latent feature, which is flattened, then followed by an MLP to reduce the 3 time dimensions, maintaining numerical stability while reducing the time dimension informatively.

For each patient, the Missingness and Time-aware Causal Language Model (MT-CLM) objective samples a random window of size m ($2 \leq m \leq M$), M being the patient's maximum time series length. To mitigate the impact of missingness, we predict only the filled portion of the m^{th} time point feature values (predicted $\hat{\mathbf{x}}_m^{(i)}$ vs ground truth $\mathbf{x}_m^{(i)}$) given first $m-1$ time points' normalized feature values and all m time stamps, minimizing:

$$\mathcal{L}_{\text{MT-CLM}} = \mathbb{E}_{i,m}\left[\left\|\mathbf{1}^{(i)}_{\text{filled},m} \odot \hat{\mathbf{x}}_m^{(i)} - \mathbf{1}^{(i)}_{\text{filled},m} \odot \mathbf{x}_m^{(i)}\right\|_2^2\right]$$

Empirically, only the time-aware CLM objective converges while MLM [36] and time-naive CLM objectives did not. This necessitates the inclusion of time stamps as a SPARC feature. The pretrained model is then finetuned with a binary cross entropy loss for each outcome (details: 4.2, A.3).

3.2 Preprocessing and featurization

We define the 'typical' feature set as the most commonly prescribed tests, selected to train the models in this work, while the 'all' set also includes less common tests (see A.4). Test values are batch-corrected given available batch information, then z-score normalized before input to the model.

4 Experiments

4.1 Datasets

The model was developed and tested with a pan-cancer dataset collected up to March 15, 2023 (*MSK-dev*), randomly split into 11 folds, with the 11$^{\text{th}}$ held out as test set and the rest for 10-fold cross validation (2719/52281 patients). Additional test data includes measurements from a patient cohort accrued from March 15, 2023 to Jan. 31, 2025 (*MSK-ts*) for prospective validation, and the cancer subsets of *MIMIC-IV-hosp* and *EHRSHOT* datasets for external validation [11, 37]. Due to limited cohort sizes and labels, only a subset of tasks are evaluated on external datasets. This aggregated dataset represents the largest cancer time series dataset for modeling to-date (details in A.4).

4.2 Baselines, experimental setups, and evaluation

Baselines and experimental setups We evaluated SPARC, SPARC-P (SPARC with MT-CLM pretraining) against diverse baselines, including statistical and deep learning models, non-temporal and temporal ones: random forest, XGBoost, and MLP trained on last available time point data, as well as random forest, XGBoost, RNN, DuETT, RainDrop, SeFT trained on temporal data (equivalent to SPARC inputs) [38–41, 27–29]. To ensure fairness, all benchmarked models are hyperparameter tuned on the first fold for each task prior to training, while SPARC is hyperparameter tuned only on the pretraining objective on the same fold (further details in A.3).

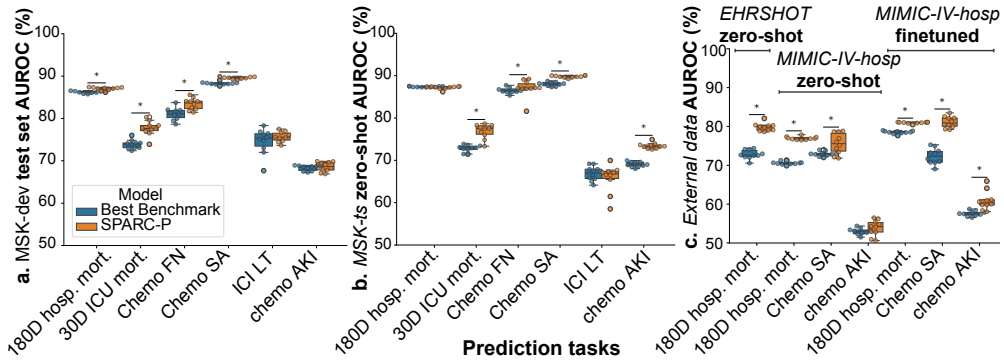


Figure 2: SPARC and benchmark performances across tasks/datasets (*: p<0.05).

Evaluation We designed prediction tasks tailored to cancer decision-making to optimize disease-specific utility (Figure 1a), [5, 6, 10–12, 26, 42–46]. Here we focus on the following oncologic tasks:
(1) 6-month mortality risk [47] after hospital discharge, used for hospice decisions.
(2) 30-day mortality risk after ICU admission, for patient-specific evaluations of ICU effectiveness.
(3-5) Febrile neutropenia (FN), severe anemia (SA), acute kidney injury within 2 weeks, after the start of high-risk chemotherapies (see details in A.1).
(6) Liver toxicity (LT) within 2 weeks, after start of high-risk immunocheckpoint inhibitors (A.1).

For all evaluated outcomes (binary classification), AUROC and AUPRC scores were calculated and compared across models, with p-values computed using a two-sided Mann-Whitney U test. We also evaluated whether the feature importance of SPARC models agree with literature expectations and identified potential novel biomarkers using LIME [48].

5 Results

SPARC-P outperformed the next-best benchmark on the internal *MSK-dev* test set at predicting 6-month post-hospitalization mortality, 30-day ICU mortality, and FN, SA, and AKI following cytotoxic chemotherapy, and performed as well as previous methods at predicting LT post-ICI treatment (Figure 4a). To assess model generalizability, we performed zero-shot inference on prospective *MSK-ts* where SPARC-P again outperformed prior methods (Figure 4b, c-left). In the *EHRSHOT* and *MIMIC-IV-hosp*, SPARC-P again outperformed prior methods, though performance declined, likely due to differences in feature space and data distribution. In *MIMIC-IV-hosp*, the large cohort size enabled partial recovery through finetuning from the pretrained SPARC model (Figure 2c, right).

5.1 Ablation studies

Benefits of pretraining Pretraining may improve prediction especially in data-scarce regimes. We investigated this by training "titrated" models with a portion of the training data (Figure 3, details in A.4). SPARC-P consistently outperforms the directly supervised SPARC in *MSK-dev* test set, demonstrating the benefits of pretraining.

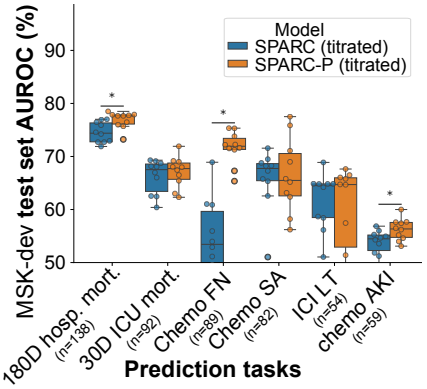


Figure 3: "Titrated" SPARC model performances with or without MT-CLM pretraining (*: p<0.05, n: train set size after titration, in random fold 1).

Feature importance analyses Historical trajectories of tests used to diagnose specific treatment complications, such as low blood hemoglobin (HgB) levels for anemia, were important and in consistent directions according to LIME (Figure 4, A.6). For longer-term prediction tasks (e.g. post hospital 180 day mortality), the model attends to more time points (A.6). We evaluated feature importance on the mortality prediction tasks which lack *a priori* hypothesized important features: across longer-term hospital discharge and shorter-term ICU admission mortality, high BUN (Blood Urea Nitrogen) consistently signified worse outcomes, identifying an important yet potentially overlooked outcome marker.

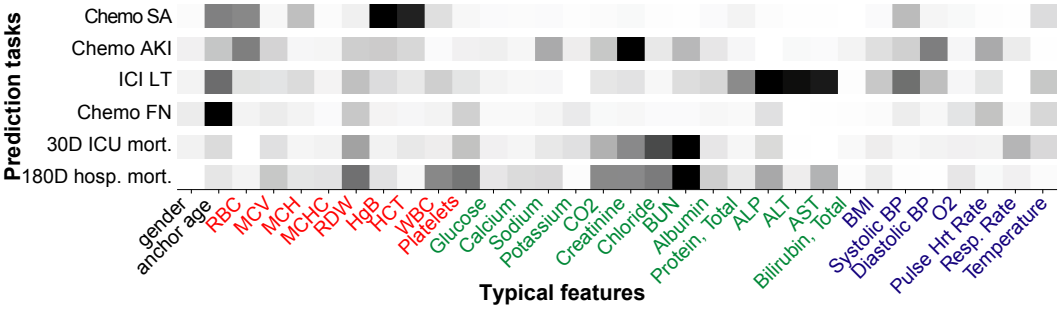


Figure 4: Averaged absolute feature importance in *MSK-dev* test set. Darker is more important.

6 Conclusion

Motivated by the divergence between ICU and cancer time series data, we introduce SPARC, a cancer outcome predictor that excels across clinically actionable prediction tasks and diverse validation scenarios. SPARC pretraining enables improved performance even in circumstances with scarce training data. SPARC predicts treatment complications using features that agree with prior expectations, and identifies putative novel mortality markers, pending further clinical validations.
Future works include validation on more external datasets of diverse patient populations, performance comparisons on clinically meaningful subcohorts, including cancer types and treatment groups, and performance analyses over more fine-grained titrations and across diverse prediction time horizons.

References

[1] Christopher R. Manz, Yichen Zhang, Kan Chen, Qi Long, Dylan S. Small, Chalanda N. Evans, Corey Chivers, Susan H. Regli, C. William Hanson, Justin E. Bekelman, Jennifer Braun, Charles A. L. Rareshide, Nina O'Connor, Pallavi Kumar, Lynn M. Schuchter, Lawrence N. Shulman, Mitesh S. Patel, and Ravi B. Parikh. Long-term effect of machine learning–triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer: A randomized clinical trial. *JAMA Oncology*, 9(3):414–418, 2023. doi: 10.1001/jamaoncol.2022.6303. URL `https://jamanetwork.com/journals/jamaoncology/fullarticle/2800545`.

[2] D. Janet Pavlin, Suzanne E. Rapp, Nayak L. Polissar, Judith A. Malmgren, Meagan Koerschgen, and Heidi Keyes. Factors affecting discharge time in adult outpatients. *Ambulatory Anesthesia*, 87:816–826, 1998.

[3] Antonia Koutsoukou. Admission of critically ill patients with cancer to the icu: many uncertainties remain. *ESMO Open*, 2(4):e000105, 2017. doi: 10.1136/esmoopen-2016-000105.

[4] Keri L. Rodriguez, Amber E. Barnato, and Robert M. Arnold. Perceptions and utilization of palliative care services in acute care hospitals. *Journal of Palliative Medicine*, 10(1):99–110, 2007. doi: 10.1089/jpm.2006.0155.

[5] Nicole M. Kuderer, Aakash Desai, Maryam B. Lustberg, and Gary H. Lyman. Mitigating acute chemotherapy-associated adverse events in patients with cancer. *Nature Reviews Clinical Oncology*, 19(11):681–697, November 2022. doi: 10.1038/s41571-022-00685-3.

[6] Filipe Martins, Latifyan Sofiya, Gerasimos P. Sykiotis, Faiza Lamine, Michel Maillard, Montserrat Fraga, Keyvan Shabafrouz, Camillo Ribi, Anne Cairoli, Yan Guex-Crosier, Thierry Kuntzer, Olivier Michielin, Solange Peters, Georges Coukos, François Spertini, John A. Thompson, and Michel Obeid. Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. *Nature Reviews Clinical Oncology*, 16(9):563–580, September 2019. doi: 10.1038/s41571-019-0218-0.

[7] Abigail E. Whitlock, Gondy Leroy, Fariba M. Donovan, and John N. Galgiani. Icd codes are insufficient to create datasets for machine learning: An evaluation using all of us data for coccidioidomycosis and myocardial infarction. *arXiv preprint arXiv:2407.07997*, 2024.

[8] Manuel Burger, Fedor Sergeev, Malte Londschien, Daphné Chopard, Hugo Yèche, Eike Gerdes, Polina Leshetkina, Alexander Morgenroth, Zeynep Babür, Jasmina Bogojeska, Martin Faltys, Rita Kuznetsova, and Gunnar Rätsch. Towards foundation models for critical care time series. *arXiv preprint arXiv:2411.16346*, 2024. doi: 10.48550/arXiv.2411.16346. URL `https://arxiv.org/abs/2411.16346`.

[9] Manuel Burger, Daphné Chopard, Malte Londschien, Fedor Sergeev, Hugo Yèche, Eike Gerdes, Polina Leshetkina, Alexander Morgenroth, Zeynep Babür, Jasmina Bogojeska, Martin Faltys, Rita Kuznetsova, and Gunnar Rätsch. A foundation model for intensive care: Unlocking generalization across tasks and domains at scale. *medRxiv preprint*, 2025. doi: 10.1101/2025.07.25.25331635. URL `https://www.medrxiv.org/content/10.1101/2025.07.25.25331635v1`.

[10] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[11] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Abigail Gayles, Ahmed Shammout, Steven Horng, Tom J. Pollard, Siqi Hao, Benjamin Moody, Brian Gow, Adarsh Kenny, George B. Moody, Sunjay Chaudhry, Roger G. Mark, Leo Anthony Celi, and Mikhail Dziadzko. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[12] U.S. Department of Health and Human Services · National Cancer Institute. *Common Terminology Criteria for Adverse Events (CTCAE) Version 5.0*. U.S. Department of Health and Human Services / National Cancer Institute, November 2017. URL `https://dctd.cancer.gov/research/ctep-trials/for-sites/adverse-events/ctcae-v5-5x7.pdf`. Published November 27, 2017.

[13] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1): 135, 2023. doi: 10.1038/s41746-023-00879-8.

[14] Chao Pang, Vincent Jeanselme, Young Sang Choi, Xinzhuo Jiang, Zilin Jing, Aparajita Kashyap, Yuta Kobayashi, Yanwei Li, Florent Pollet, Karthik Natarajan, and Shalmali Joshi. Fomoh: A clinically meaningful foundation model evaluation for structured electronic health records. *arXiv preprint arXiv:2505.16941*, 2025.

[15] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020.

[16] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1):86, 2021.

[17] Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021.

[18] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*, 2021. URL `https://arxiv.org/abs/2107.03134`.

[19] Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.

[20] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, et al. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194, 2022. doi: 10.1038/s41746-022-00742-2.

[21] Zeljko Kraljevic, Daniel Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

[22] Ethan Steinberg, Yizhe Xu, Jason Alan Fries, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NialiwI2V6`.

[23] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehr data. In *The Thirteenth International Conference on Learning Representations*, 2024. URL `https://arxiv.org/abs/2412.16178`.

[24] Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ Digital Medicine*, 7(1):256, 2024.

[25] Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, SeongEun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. Genhpf: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, 2023.

[26] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[27] Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G. Krishnan. Duett: Dual event time transformer for electronic health records. *arXiv preprint arXiv:2304.13017*, 2023.

[28] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Raindrop: Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations (ICLR)*, 2022.

[29] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. *arXiv preprint arXiv:1909.12064*, 2019.

[30] Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *arXiv preprint arXiv:2107.14293*, 2021.

[31] Tian-Gen Chang, Yingying Cao, Hannah J. Sfreddo, Saugato Rahman Dhruba, Se-Hoon Lee, Cristina Valero, Seong-Keun Yoo, Diego Chowell, Luc G. T. Morris, and Eytan Ruppin. Loris robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features. *Nature Cancer*, 5(8):1158–1175, 2024.

[32] Seong-Keun Yoo, Conall W. Fitzgerald, Byuri Angela Cho, Bailey G. Fitzgerald, Catherine Han, Elizabeth S. Koh, Abhinav Pandey, Hannah Sfreddo, Fionnuala Crowley, Michelle Rudshteyn Korostin, Neha Debnath, Yan Leyfman, Cristina Valero, Mark Lee, Joris L. Vos, Andrew Sangho Lee, Karena Zhao, Stanley Lam, Ezekiel Olumuyide, Fengshen Kuo, Eric A. Wilson, Pauline Hamon, Clotilde Hennequin, Miriam Saffern, Lynda Vuong, A. Ari Hakimi, Brian Brown, Miriam Merad, Sacha Gnjatic, Nina Bhardwaj, Matthew D. Galsky, Eric E. Schadt, Robert M. Samstein, Thomas U. Marron, Mithat Gönen, Luc G. T. Morris, and Diego Chowell. Prediction of checkpoint inhibitor immunotherapy efficacy for cancer using routine blood tests and clinical data. *Nature Medicine*, 31(3):869–880, 2025.

[33] Simon Mantha, Subrata Chatterjee, Rohan Singh, John Cadley, Chester Poon, Avijit Chatterjee, Daniel Kelly, Michelle Sterpi, Gerald Soff, Jeffrey Zwicker, José Soria, Magdalena Ruiz, Andres Muñoz, and Maria Arcila. Application of machine learning to the prediction of cancer-associated venous thromboembolism. *Research Square preprint*, 2023. doi: 10.21203/rs.3.rs-2870367/v1. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10197737/`.

[34] Jiheum Park, Michael G. Artin, Kate E. Lee, Yoanna S. Pumpalova, Myles A. Ingram, Benjamin L. May, Michael Park, Chin Hur, and Nicholas P. Tatonetti. Deep learning on time series laboratory test results from electronic health records for early detection of pancreatic cancer. *Journal of Biomedical Informatics*, 131:104095, 2022. doi: 10.1016/j.jbi.2022.104095. URL `https://pubmed.ncbi.nlm.nih.gov/35598881/`.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. doi: 10.48550/arXiv.1706.03762. URL `https://arxiv.org/abs/1706.03762`.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. doi: 10.48550/arXiv.1810.04805. URL `https://arxiv.org/abs/1810.04805`.

[37] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A Fries, and Nigam H Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *arXiv preprint arXiv:2307.02028*, 2023. URL `https://arxiv.org/abs/2307.02028`.

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

[39] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016. URL `https://arxiv.org/abs/1603.02754`.

[40] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5):183–197, 1991. doi: 10.1016/0925-2312(91)90023-5. URL `https://doi.org/10.1016/0925-2312(91)90023-5`.

[41] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[42] Carelon Medical Benefits Management. Febrile neutropenia risk. https://guidelines.carelonmedicalbenefitsmanagement.com/febrile-neutropenia-risk/, 2024. Effective Date: 2024-02-01; Last Review Date: 2023-10-23.

[43] Authors from the article—please supply once known. [title from the sciencedirect article—please supply once known]. *Journal Name*, Volume(Issue):Page range, 2022. doi: 10.1016/S0305-7372(22)00091-3.

[44] Pieter Braet, G. V. R. Sartò, M. Pirovano, B. Sprangers, and L. Cosmai. Treatment of acute kidney injury in cancer patients. *Clinical Kidney Journal*, 15(5):873–884, 2021. doi: 10.1093/ckj/sfab292.

[45] Saint Luke's Health System. Cancer treatment and kidney damage (nephrotoxicity). https://www.saintlukeskc.org/health-library/cancer-treatment-and-kidney-damage-nephrotoxicity. Accessed: 2025-08-11.

[46] Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. Section 2: Aki definition. *Kidney International Supplements*, 2(1):19–36, 2012. doi: 10.1038/kisup.2011.32. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4089595/.

[47] National Institute on Aging. What are palliative care and hospice care? https://www.nia.nih.gov/health/hospice-and-palliative-care/what-are-palliative-care-and-hospice-care, 2021. Accessed: 2025-08-11.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *arXiv preprint arXiv:1602.04938*, 2016. doi: 10.48550/arXiv.1602.04938. URL https://arxiv.org/abs/1602.04938.

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. URL https://arxiv.org/abs/1711.05101. Submitted Nov 14, 2017; Revised Jan 4, 2019.

[50] Donavan T Cheng, Talia N Mitchell, Ahmet Zehir, Ronak H Shah, Ryma Benayed, Aijazuddin Syed, Raghu Chandramohan, Zhen Yu Liu, Helen H Won, Sasinya N Scott, et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (msk-impact): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *Journal of Molecular Diagnostics*, 17(3):251–264, 2015. doi: 10.1016/j.jmoldx.2014.12.006.

7 Appendix

7.1 Appendix A.1

Table A.1 Description of prediction tasks. For treatment complication tasks, only patients taking medications known to elevate risks of each complication (listed) are included. Note that additionally, for treatment complication tasks, in addition to these task-specific inclusion criteria, a patient has to have at least 3 historical and 3 future (post-cutoff) corresponding labs to be included to reduce the risk of erroneous labeling of positive/negative patients. For each of these tasks, data is included up to a two-week prediction horizon prior to positive symptom onset, up to treatment stop.

Prediction task	Included medications	Positive criteria	Inclusion criterion
Mortality within 180 days of hospital discharge	pan-medication (180D hosp. mort.)	Death within 180 days	Death within or last contact date post 180 days of hospital discharge (censored patients excluded)
Mortality within 30 days of ICU admission (30D ICU mort.)	pan-medication	Death within 30 days	Death within or last contact date post 30 days of ICU admission (censored patients excluded)
severe anemia (SA)	carboplatin, etoposide, docetaxel, doxorubicin, cisplatin, gemcitabine	HgB<=8	No historical anemia
febrile neutropenia (FN)	carboplatin, etoposide, docetaxel, doxorubicin, cisplatin, gemcitabine	Neutrophil nount<1000 AND temperature>=38°C	No historical febrile neutropenia
Acute kidney injury (AKI)	cisplatin, ifosfamide, methotrexate, pemetrexed, cyclophosphamide	An at least 0.3mg/dl increase in Creatinine within 2 days OR an at least 1.5X elevation from baseline (the patient's average of first two Creatinine values)	No historical AKI
Liver toxicity (LT)	ipilimumab, atezolizumab, nivolumab, pembrolizumab	Grade 3 and above sinusoidal obstruction syndrome (Bilirubin >5mg/dL) OR grade 3 and above increase in liver enzymes AST, ALT, and ALP (>5 folds from baseline, calculated from the patient's average of historical lab measurements)	No historical liver toxicity, with baseline of AST, ALT, ALP tests calculated from the maximum of 33U/L, 36U/L, 147IU/L and the patient's average of their first two respective test measurements.

7.2 Appendix A.2

Table A.2 Features and descriptions. “Typical" features are highly prevalent as they are typically prescribed together and performed at nearly every check-up. Others are prescribed less frequently and at discretion, such as concerns about inflammation (immune cells), specific cancer types (cancer antigen tests), or during intensive care (invasive blood pressure). Gender and anchoring age (see A.4 are also included.

Test	Unit	Lab panel/vitals	Feature set
Gender	NA	Static	typical, all
Anchoring age	NA	Static	typical, all
RBC	Million/mcL	Complete Blood Count	typical, all
MCV	fL	Complete Blood Count	typical, all
MCH	pg	Complete Blood Count	typical, all
MCHC	g/dL	Complete Blood Count	typical, all
RDW	%	Complete Blood Count	typical, all
HgB	g/dL	Complete Blood Count	typical, all
HCT	%	Complete Blood Count	typical, all
WBC	K/mcL	Complete Blood Count	typical, all
Platelets	K/mcL	Complete Blood Count	typical, all
Glucose	mg/dL	Comprehensive Metabolic Panel	typical, all
Calcium	mg/dL	Comprehensive Metabolic Panel	typical, all
Sodium	mEq/L	Comprehensive Metabolic Panel	typical, all
Potassium	mEq/L	Comprehensive Metabolic Panel	typical, all
CO2	mEq/L	Comprehensive Metabolic Panel	typical, all
Creatinine	mg/dL	Comprehensive Metabolic Panel	typical, all
Chloride	mEq/L	Comprehensive Metabolic Panel	typical, all
BUN	mg/dL	Comprehensive Metabolic Panel	typical, all
Albumin	g/dL	Comprehensive Metabolic Panel	typical, all
Protein, Total	g/dL	Comprehensive Metabolic Panel	typical, all
ALP	U/L	Comprehensive Metabolic Panel	typical, all
ALT	U/L	Comprehensive Metabolic Panel	typical, all
AST	U/L	Comprehensive Metabolic Panel	typical, all
Bilirubin, Total	mg/dL	Comprehensive Metabolic Panel	typical, all
Conj. Bilirubin	mg/dL	Comp. Metabolic Panel & Conj Bilirubin	all
Eosinophils	%	Complete Blood Count (CBC), With Differential	all
Immature Granulocyte	mg/dL	Complete Blood Count (CBC), with Differential	all
Leucocytes	%	Complete Blood Count (CBC), with Differential	all
Lymphocytes	%	Complete Blood Count (CBC), with Differential	all
Variant lymphocytes	%	Complete Blood Count (CBC), with Differential	all
Megakaryocyte Fragment	/100 WBC	Complete Blood Count (CBC), with Differential	all
Metamyelocytes	%	Complete Blood Count (CBC), with Differential	all
Monocytes	%	Complete Blood Count (CBC), with Differential	all
Myelocytes	%	Complete Blood Count (CBC), with Differential	all
Neutrophils	%	Complete Blood Count (CBC), with Differential	all
Nucleated RBC	/100 WBC	Complete Blood Count (CBC), with Differential	all
Promyelocytes	%	Complete Blood Count (CBC), with Differential	all
Basophils	%	Complete Blood Count (CBC), with Differential	all
Blasts	%	Complete Blood Count (CBC), with Differential	all
PSA	ng/mL	Cancer Antigen Test	all
TSH	mIU/L	Cancer Antigen Test	all
Alphafetoprotein	ng/mL	Cancer Antigen Test	all
B2M	mg/L	Cancer Antigen Test	all
CEA	ng/mL	Cancer Antigen Test	all
Cancer Antigen 125	U/mL	Cancer Antigen Test	all
Cancer Antigen 15-3	U/mL	Cancer Antigen Test	all
Cancer Antigen 19-9	U/mL	Cancer Antigen Test	all
BMI	kg/m^2	Vitals	typical, all
Systolic Blood Pressure (BP)	mmHg	Vitals	typical, all
Diastolic Blood Pressure (BP)	mmHg	Vitals	typical, all
Invasive systolic blood pressure	mmHg	Vitals	all
Invasive diastolic blood pressure	mmHg	Vitals	all
O2	%	Vitals	typical, all
Pulse Heart Rate	bpm	Vitals	typical, all
Resp. rate	breaths pm	Vitals	typical, all
Temperature	°C	Vitals	typical, all

A.3 Appendix A.3

Further details on SPARC and benchmark implementation and training details.

Grid-search hyperparameter tuning For each of the benchmarked (non-SPARC) models for each task, hyperparameter tuning was done using the first fold out of 10 non-heldout fold to achieve the highest validation AUPRC. For deep learning models, a diverse range of parameters were tested to identify the best performing combination, including learning rate, weight decay, dropout (specific to DuETT), number of transformer encoder layers (specific to DuETT), maximum number of time series included (specific to seFT). For SPARC, a non-task specific hyperparameter tuning is conducted based on minimizing the first-fold validation loss calculated with the MT-CLM pretraining reconstruction loss, where all the above-mentioned hyperparameters are tuned. For statistical learning models (random forest and xgBoost), a grid search over hyperparameters 'n_estimaters' and 'max_depth' (and for random forest additionally 'min_samples_split' and 'min_samples_leaf' is done every time such a model is trained - no prior hyperparameter necessary.

Model implementation details For all the deep learning models, an AdamW optimizer [49] specifying the learning rate and weight decay from hyperparameter tuning experiments is used for gradient descent, with a learning rate scheduler ReduceLROnPlateau to further smooth changes in learning rate. Codes (unanonymized) will be provided upon request.

A.4 Appendix A.4

Table A.4 Datasets and descriptions. To support the sustained monitoring of treatment complication risks and to predict the mortality risks post ICU admission, data prior to ICU admission, especially those from long-term care (at least multiple dates) is required. Data from solely ICU stays are not suitable for these tasks. We cleaned the following datasets to acquire the features as described in A.1

. Importantly, the first measurement for each lab test per unique date is included to avoid duplicates.

Dataset	Institution	Feature extracted	# cancer patients	(25, 50, 75)% # dates	# unique extracted entries
MSK-dev	MSKCC	typical, all	57,510	(20, 41, 77)	Labs: 70,666,915 vitals: 8,324,608
MSK-ts	MSKCC	typical, all	15,092	(9, 18, 32)	Labs: 7,825,570 vitals: 924,216
EHRSHOT (cancer subset)	Stanford Health	typical	544	(28, 58.5, 108)	Labs: 442,326 vitals: 204,083
MIMIC-IV-hosp (cancer subset)	Beth Israel Deaconess MC	typical	26,969	(10, 24, 52)	Labs: 16,035,591 vitals: 0 (not included)

Further details regarding the above cleaned datasets:
1. To ensure the quality of training data, each patient included in *MSK-dev* has to have one of each of the 23 laboratory tests in the 'typical' set.
2. For the *MIMIC-IV-hosp* dataset, vitals are not included per official documentations [11]. Although occasional vitals data does exist in the data, they exist at way lower abundance compared to labs and are not inducive to constructing the febrile neutropenia prediction task.
3. Anchoring age is the age at which each patient in the cohort is at their cohort-specific anchoring time. For *MSK-dev, MSK-ts*, the patients are all subsets of the *MSK-IMPACT*[50] targeted-sequencing cohort, and thus the anchoring age is the age at the first sequencing report. For *MIMIC-IV-hosp, EHRSHOT*, we used the age at the first collected lab/vital test as the anchoring age.
4. For titration experiments, $\frac{1}{125}$ of the training set is used for all tasks except for the 30-day post ICU admission mortality risk prediction task due to the limited number of training samples. For the ICU task (2384 training samples from *MSK-dev*), we use $\frac{1}{25}$ of the training set for titration.

A.5 Appendix A.5

Table A.5.1 Model performance on *MSK-dev* test set

Model	Metric	180D hosp. mortality	30D ICU mortality	Chemo SA	Chemo AKI	ICI LT	Chemo FN
lasttp_MLP	AUROC	81.140±0.25	73.460±0.82	73.717±11.90	50.614±0.84	56.206±5.08	50.000±0.00
SeFT	AUROC	73.115±2.95	61.347±2.71	71.028±0.18	62.647±0.22	64.054±0.96	77.670±0.48
Raindrop	AUROC	69.314±0.31	62.818±4.20	65.857±2.33	56.474±0.26	60.020±1.77	74.601±1.78
RNN	AUROC	86.249±0.28	72.765±1.26	88.271±0.33	68.325±0.61	74.597±2.89	79.297±1.13
lasttp_rf	AUROC	72.545±0.60	66.360±1.11	77.152±0.70	51.249±1.18	52.530±0.87	50.348±1.04
lasttp_xgboost	AUROC	73.701±0.81	66.814±1.82	76.645±0.81	56.382±1.34	55.567±1.31	50.500±0.57
temporal_rf	AUROC	70.727±1.49	68.170±1.13	76.495±0.71	52.440±1.00	50.435±0.69	50.600±0.77
temporal_xgboost	AUROC	76.770±0.65	68.623±1.52	79.380±0.74	55.522±0.61	55.111±1.48	51.319±1.34
DuETT	AUROC	83.693±0.14	73.874±1.00	80.583±0.28	65.180±0.96	63.929±2.10	81.063±1.39
SPARC	AUROC	86.956±0.20	77.087±1.12	89.852±0.19	**68.690±1.56**	76.259±1.43	82.313±2.58
SPARC-P	AUROC	**87.013±0.38**	**77.668±1.66**	89.603±0.26	**68.596±1.08**	**75.659±1.14**	**83.407±1.31**
lasttp_MLP	AUPRC	87.221±0.24	72.797±0.52	62.828±14.85	23.797±0.39	18.565±5.72	4.234±0.00
SeFT	AUPRC	79.461±1.91	58.345±2.24	59.637±0.39	34.872±0.59	22.811±1.18	24.325±1.11
Raindrop	AUPRC	78.093±0.24	57.885±4.28	49.979±2.26	28.609±0.36	18.964±1.73	17.460±1.57
RNN	AUPRC	90.979±0.16	71.123±2.09	80.273±0.69	42.986±1.06	32.442±4.16	25.170±1.08
lasttp_rf	AUPRC	75.214±0.47	61.059±0.97	61.703±0.75	23.886±1.07	14.664±1.36	4.567±1.00
lasttp_xgboost	AUPRC	76.235±0.64	60.856±1.66	60.355±0.92	27.515±1.36	17.384±1.94	4.547±0.36
temporal_rf	AUPRC	73.690±1.07	62.220±1.14	61.108±0.81	25.488±1.00	12.073±1.13	4.747±0.65
temporal_xgboost	AUPRC	79.845±0.47	62.566±1.43	64.387±1.10	27.404±0.67	17.807±1.87	5.883±1.90
DuETT	AUPRC	89.046±0.12	69.577±1.24	69.837±0.70	36.906±1.05	21.007±1.87	26.764±1.86
SPARC	AUPRC	91.692±0.15	**75.607±0.83**	**81.693±0.57**	**43.844±2.24**	**38.776±2.11**	28.241±2.42
SPARC-P	AUPRC	**91.793±0.23**	75.556±1.84	81.567±0.54	43.412±1.42	39.362±1.12	**29.812±1.65**

Table A.5.2 Model performance on *MSK-ts* dataset (prospective zero-shot validation)

Model	Metric	180D hosp. mortality	30D ICU mortality	Chemo SA	Chemo AKI	ICI LT	Chemo FN
lasttp_MLP	AUROC	78.780±0.96	67.514±1.51	73.639±11.84	52.299±1.29	53.512±2.34	50.000±0.00
SeFT	AUROC	70.997±1.47	57.094±1.00	74.365±0.12	63.979±0.41	58.069±0.78	85.826±0.60
Raindrop	AUROC	68.448±0.25	62.905±0.90	66.907±4.43	54.397±0.50	58.184±1.10	84.589±0.18
RNN	AUROC	**87.369±0.10**	72.808±0.77	88.013±0.44	69.049±0.56	**66.885±1.63**	85.730±0.74
lasttp_rf	AUROC	73.384±0.32	64.222±1.33	77.391±0.69	55.173±0.99	50.750±0.48	50.575±1.46
lasttp_xgboost	AUROC	71.820±0.59	64.808±1.54	78.141±0.70	57.737±1.22	51.552±0.55	50.920±0.92
temporal_rf	AUROC	71.113±1.79	65.926±1.77	76.846±0.67	51.503±1.31	49.917±0.04	51.598±1.67
temporal_xgboost	AUROC	77.936±0.47	65.804±1.58	80.525±0.52	57.818±0.79	51.343±0.70	52.179±1.15
DuETT	AUROC	83.999±0.08	72.805±1.06	83.059±0.46	60.593±1.84	61.374±1.24	86.513±0.63
SPARC	AUROC	87.102±0.23	75.769±1.02	89.413±0.23	72.649±2.08	65.063±1.25	87.336±1.55
SPARC-P	AUROC	87.160±0.37	**76.870±1.62**	**89.778±0.24**	**73.398±0.64**	65.784±3.16	**87.035±2.01**
lasttp_MLP	AUPRC	69.971±1.49	73.539±1.87	60.409±15.09	22.106±1.40	11.480±1.89	2.696±0.00
SeFT	AUPRC	53.364±1.82	61.394±0.82	59.502±0.30	33.517±0.39	15.753±0.84	28.023±0.58
Raindrop	AUPRC	52.758±0.40	68.095±0.63	50.571±3.38	22.396±0.26	14.015±1.02	23.500±1.37
RNN	AUPRC	**78.709±0.26**	77.614±0.82	76.011±0.77	40.614±1.01	17.885±1.15	27.549±1.36
lasttp_rf	AUPRC	48.976±0.40	67.650±0.98	56.156±0.60	24.140±0.82	8.647±0.33	3.181±1.03
lasttp_xgboost	AUPRC	47.316±0.62	67.597±1.08	57.084±0.75	25.089±1.12	8.948±0.36	3.251±0.72
temporal_rf	AUPRC	46.459±1.74	68.444±1.30	56.266±0.72	20.875±1.12	8.207±0.00	3.824±1.35
temporal_xgboost	AUPRC	57.334±0.60	68.344±1.17	60.529±0.69	26.001±0.95	9.362±1.00	4.945±1.36
DuETT	AUPRC	73.583±0.21	77.744±1.28	67.915±1.03	28.389±1.97	15.378±1.18	31.115±1.19
SPARC	AUPRC	77.627±0.41	79.670±1.19	78.250±0.92	41.042±4.29	18.612±1.75	**34.789±3.17**
SPARC-P	AUPRC	78.515±0.65	**80.697±0.98**	**79.671±0.58**	**41.142±1.85**	**19.173±1.91**	32.295±5.27

Table A.5.3 MSK-trained model zeroshot generalization performance on *EHRSHOT, MIMICIV-hosp*

Model	Metric	180D hosp. mort EHRSHOT	180D hosp. mort MIMICIV_hosp	Chemo SA MIMICIV_hosp	Chemo AKI MIMICIV_hosp
lasttp_MLP	AUROC	65.386±2.55	66.767±1.01	68.795±9.56	50.739±1.03
SeFT	AUROC	69.111±6.10	62.175±0.85	55.600±0.36	45.314±0.39
Raindrop	AUROC	63.669±0.92	58.716±0.77	54.918±1.89	50.900±0.52
RNN	AUROC	48.282±1.53	62.088±1.49	66.013±1.08	40.953±3.10
lasttp_rf	AUROC	67.294±1.97	67.675±0.28	64.747±1.67	50.965±0.77
lasttp_xgboost	AUROC	67.066±1.90	65.219±0.32	65.253±1.54	52.251±1.56
temporal_rf	AUROC	69.509±1.33	67.439±1.02	72.937±0.58	51.175±0.63
temporal_xgboost	AUROC	71.339±1.61	63.859±1.14	72.067±1.04	52.632±1.36
DuETT	AUROC	72.803±1.10	70.587±0.40	68.776±1.46	51.766±0.84
SPARC	AUROC	77.961±1.51	76.967±0.47	70.798±2.60	**52.890±1.57**
SPARC-P	AUROC	**79.815±0.96**	**76.968±0.40**	**75.596±2.52**	53.857±1.93
lasttp_MLP	AUPRC	68.371±1.99	58.695±1.34	53.615±9.95	40.140±1.14
SeFT	AUPRC	67.590±5.86	46.418±0.72	37.676±0.38	36.856±0.63
Raindrop	AUPRC	67.243±0.87	45.776±0.53	37.779±1.36	38.471±0.40
RNN	AUPRC	51.312±1.65	46.301±1.42	48.439±1.07	33.323±1.27
lasttp_rf	AUPRC	62.314±1.39	49.538±0.42	48.634±1.13	40.093±0.89
lasttp_xgboost	AUPRC	61.690±1.56	46.728±0.33	48.243±1.74	40.733±1.41
temporal_rf	AUPRC	63.400±1.14	48.184±1.18	54.209±0.88	40.262±0.86
temporal_xgboost	AUPRC	66.475±1.65	48.704±0.74	52.239±1.27	40.702±0.92
DuETT	AUPRC	73.411±1.01	60.164±0.44	52.811±2.09	40.523±1.08
SPARC	AUPRC	77.107±1.67	66.470±0.50	56.265±3.69	43.549±1.60
SPARC-P	AUPRC	**80.976±1.01**	**66.942±0.45**	**62.682±3.42**	**43.027±1.84**

Table A.5.4 Finetuned model performance on *EHRSHOT, MIMICIV-hosp*

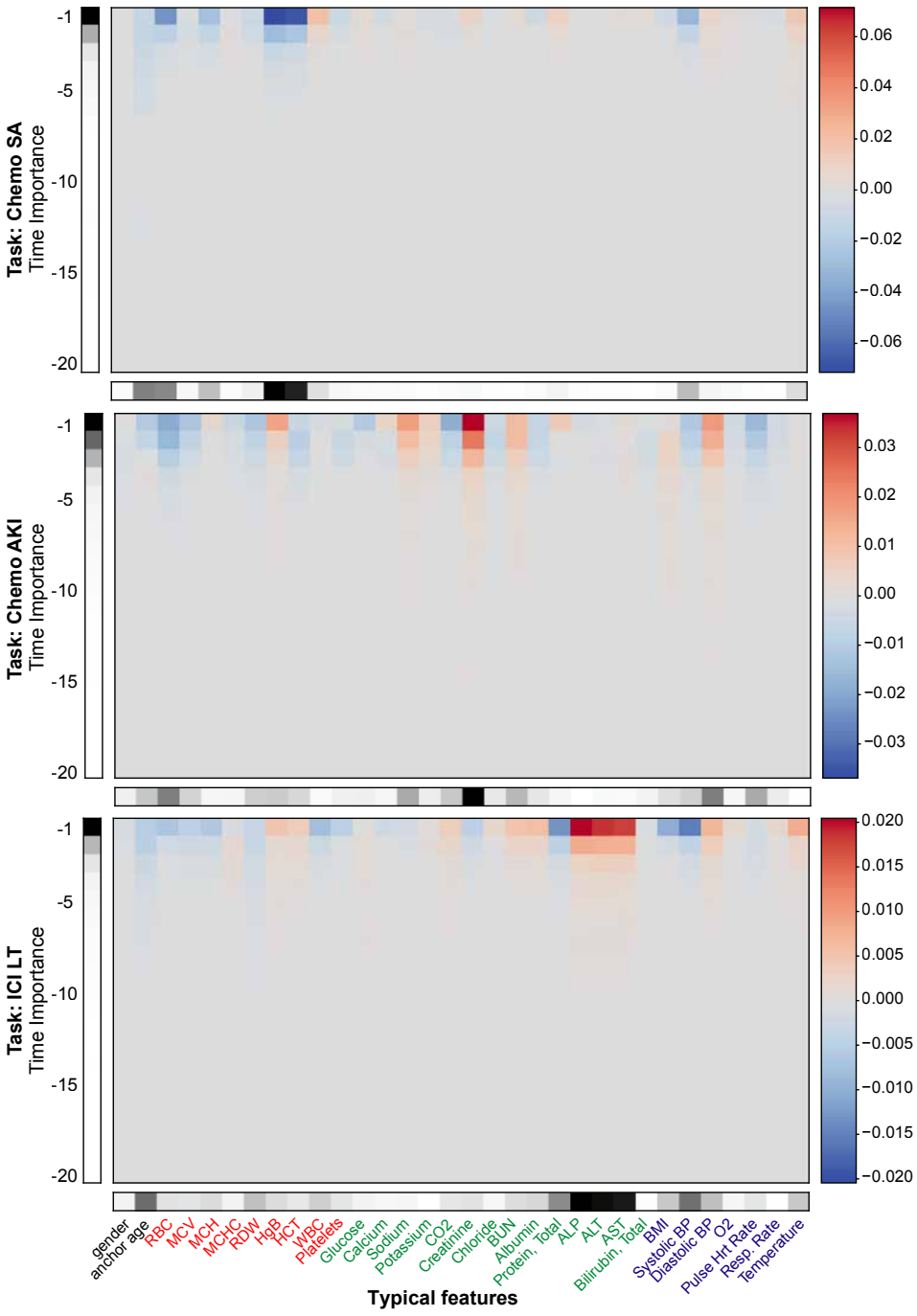
Model	Metric	180D hosp. mort MIMICIV_hosp	Chemo SA MIMICIV_hosp	Chemo AKI MIMICIV_hosp
lasttp_MLP	AUROC	68.952±0.50	60.220±12.59	49.632±0.74
SeFT	AUROC	70.234±0.15	65.482±0.18	57.577±0.64
Raindrop	AUROC	67.169±0.52	59.692±3.73	50.447±10.26
RNN	AUROC	78.489±0.38	68.009±5.24	51.146±4.80
lasttp_rf	AUROC	65.191±1.29	71.752±1.52	56.133±2.55
lasttp_xgboost	AUROC	64.674±0.00	69.172±0.00	53.619±0.00
temporal_rf	AUROC	67.060±2.00	72.326±1.92	56.660±3.24
temporal_xgboost	AUROC	54.637±0.00	72.179±0.00	55.335±0.00
DuETT	AUROC	75.546±0.65	70.790±2.74	51.882±6.48
SPARC	AUROC	80.424±0.17	79.140±1.17	53.495±4.14
SPARC-P	AUROC	**80.960±0.18**	**81.097±1.38**	**60.881±2.31**
lasttp_MLP	AUPRC	60.522±0.56	40.791±10.19	38.784±0.47
SeFT	AUPRC	57.229±0.26	47.628±0.45	44.497±0.47
Raindrop	AUPRC	54.068±0.51	41.327±4.30	41.742±7.91
RNN	AUPRC	70.292±0.51	49.552±4.99	40.957±5.14
lasttp_rf	AUPRC	46.643±2.43	51.246±2.06	42.681±1.86
lasttp_xgboost	AUPRC	45.802±0.00	48.167±0.00	40.650±0.00
temporal_rf	AUPRC	47.830±2.02	53.182±2.00	44.359±3.09
temporal_xgboost	AUPRC	39.381±0.00	50.249±0.00	41.561±0.00
DuETT	AUPRC	64.734±0.94	49.141±3.02	42.469±5.52
SPARC	AUPRC	71.467±0.28	**63.087±2.20**	42.895±4.42
SPARC-P	AUPRC	**72.585±0.35**	61.690±2.94	**51.431±1.80**

Notes:
1. Highest performing model per metric per task per dataset is labelled in **bold** and second highest labelled in underline.
2. *MSK-dev* and *MSK-timeshift* share the same, 'typical' feature space, and the age feature is anchored around first sequencing report dates for both dataset. *EHRSHOT* shares the majority of the "typical"

feature space, but due to its limited size (522 cancer patients with 183 non-censored patients for the 180D hosp. mort. task), only this task is evaluated on in a zero-shot setting - a direct deployment of the MSK-trained model.
3. *MIMICIV-hosp* dataset has an abundant number of patients, but does not contain the vitals data (temperature in particular) to define Chemo FN task, nor enough patients who were recorded to take one of the ICIs known to elevate liver toxicity risks (see A.1 for the list), and as a hospital cohort defined by ED/ICU admission, cautions need to made about collected pre-ICU data - thus only the 180D hosp. mort., Chemo SA, Chemo AKI tasks were evaluated in both a zeroshot (direct deployment of MSK models) and a finetuned setting (1/3 of the data heldout for testing, 1/3 for training and 1/3 for validation, with different initialization seeds for generality).

A.6 Appendix A.6

Feature importance analysis detailed results.



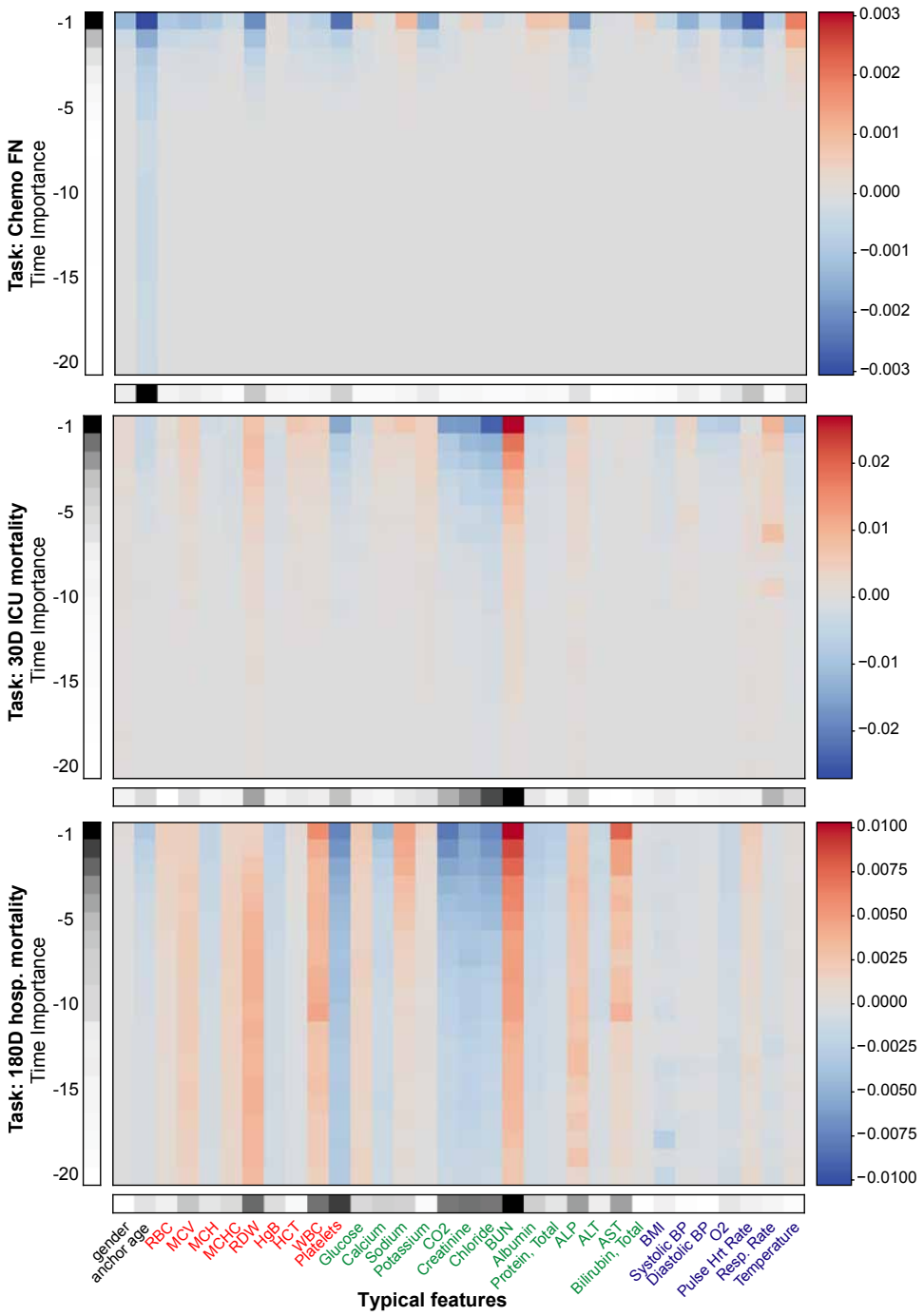


Figure A.5. Feature importance analyses over time and feature dimension, averaged across 10 folds (models) and patients in the *MSK-dev* dataset.