# Machine learning algorithm for mortality prediction in patients with advanced penile cancer

Robert Chen, PhD
Emory University School of Medicine

## ABSTRACT

Penile cancer remains a rare cancer with an annual incidence of 1 in 100,000 men in the United States, accounting for 0.4-0.6% of all malignancies. Furthermore, to date there are no predictive models of early mortality in penile cancer. Meanwhile, machine learning has potential to serve as a prognostic tool for patients with advanced disease.

We developed a machine learning model for predicting early mortality in penile cancer (survival less than 11 months after initial diagnosis. A cohort of 88 patients with advanced penile cancer was extracted from the Surveillance, Epidemiology and End Results (SEER) program. In the cohort, patients with advanced penile cancer exhibited a median overall survival of 21 months, with the 25th percentila of overall survival being 11 months. We constructed predictive features based on patient demographics, staging, metastasis, lymph node biopsy criteria, and metastatic sites. We trained a multivariate logistic regression model, tuning parameters with respect to regularization, and feature selection criteria. Upon evaluation with 5-fold cross validation, our model achieved 68.2% accuracy with AUC 0.696.

Criteria for advanced staging (T4, group stage IV), as well as higher age, white race and squamous cell histology, were the most predictive of early mortality. Tumor size was the strongest negative predictor of early mortality.

Our study showcases the first known predictive model for early mortality in patients with advanced penile cancer and should serve as a framework for approaching the clinical problem in future studies. Future work should aim to incorporate other data sources such as genomic and metabolomic data, increase patient counts, incorporate clinical characteristics such as ECOG and RECIST criteria, and assess the performance of the model in a prospective fashion.

## Keywords

penile cancer, machine learning, cancer prognosis, electronic health records, biomarkers

## 1. INTRODUCTION

Due to recent advances in diagnostic modalities such as fine–needle biopsy (FNB) and dynamic sentinel node biopsy (DSNB), stratification of patients with penile cancer has enabled clinicians to ascertain important features of penile cancer that impact the treatment course[1]. Despite this, prognosis remains poor for patients with penile cancer. Penile cancer carries a 5-year survival rate of 67%, and 12% in advanced stages[2,3]. Meanwhile, to date there does not exist a robust, interpretable prognostic algorithm for penile cancer patients with respect to mortality risk. However, it is known that there are several demographic and clinical correlates of penile cancer incidence. For example, penile cancer has higher incidence in some areas such as Asia, Africa and South America. In these regions, penile cancer accounts for approximately 10% of all malignancies among men [3]. Furthermore, there is a higher rate of penile cancer among Hispanics compared to non-Hispanics [4–6]. Such disparities in incidence of penile cancer naturally lead one to posit that there may exist clinical or demographic correlates of mortality risk between patients.

While there does not exist an intuitive method for risk stratification of patients with penile cancer, there is strong evidence of the potential of machine learning for stratification of patients in other diseases such as heart failure [7–10], kidney disease [11], and critical care [12–17]. Furthermore, machine learning has been shown to be effective for readmission prediction [17–19], drug adverse event prediction [20]. While such a method does not exist, we posit that a machine learning-based method can be useful for clinical decision support in the management of penile cancer.

In this study we developed a machine learning model based on logistic regression for prediction of early mortality in patients with penile cancer from retrospective real-world data and evaluated their performance. Our model leverages predictive features in a variety of domains including demographics, histology, staging, tumor spread and metastatic status.

## 2. METHODS

A cohort of patients was selected from the The Surveillance, Epidemiology and End Results (SEER) program public retrospective dataset[21].

### 2.1 Cohort Construction

Patients were selected from The Surveillance, Epidemiology and End Results (SEER) program was used to identify a cohort of 6,201 male patients who were diagnosed with penile cancer between 2010 and 2015. Of these patients, inclusion and exclusion criteria were applied, resulting in a cohort of 88 patients to be used in the machine learning model.

**Inclusion criteria**: Patients were included if they had the following associated ICD-9 codes for penile cancer: C600, C601, C602, C608, C609. The patient is required to have one of the following stages, corresponding to advanced penile cancer:

- any T, N1 (i.e. a palpable mobile unilateral inguinal lymph node), M0 or;

- any T, N2 (i.e. palpable mobile multiple or bilateral inguinal lymph nodes), M0 or;
- any T, N3 (i.e. fixed inguinal nodal mass or any pelvic lymphadenopathy), M0

Furthermore, patients are included if they have a date of diagnosis between 2010 and 2015.

**Exclusion criteria**: Patients are excluded if they did not have follow-up information following their initial diagnosis.

Table 1 shows descriptive statistics of the cohort.

| | **Survival < 11 mos** | **Survival >= 11 mos** | **Total** |
|---|---|---|---|
| **Characteristic** | *n*= 22 | *n* = 63 | *N*= 85 |
| **Age (mean)** | 69.4 | 65.6 | 66.7 |
| **Sex (% male)** | 100% | 100% | 100% |
| **Race (%)** | | | |
| American Indian/Asian | 0% | 6.3% | 4.7% |
| Black | 9.1% | 11.1% | 10.6% |
| White | 90.9% | 82.5% | 84.7% |
| **Group Stage (%)** | | | |
| III | 45.4% | 69.8% | 63.5% |
| IV | 54.5% | 30.2% | 36.5% |
| **Tumor Size (mean, cm)** | 2.59cm | 3.25cm | 3.08cm |
| **Histology (%)** | | | |
| Squamous cell neoplasms | 100% | 95.2% | 96.5% |
| Transitional cell papillomas and carcinomas | 0% | 1.6% | 1.2% |
| Nevi and melanomas | 0% | 3.2% | 2.4% |

**Table 1**: Baseline characteristics of the study cohort.

Features corresponding to several domains were extracted for the cohort. These include demographics, AJCC staging criteria, and metastatic sites.. Table 2 shows the description of features used in the model.

## 2.2 Kaplan Meier Analysis

A Kaplan Meier[22] analysis was performed with all patients in the cohort. The 25th percentile of overall survival time in months, was used as a threshold to determine class labels of patients:

- 0: patient survived at least the 25th percentile of overall survival
- 1: patient death occurred before the 25th percentile of overall survival

| **Feature** | **No. of features** | **Example** | **Aggregation** |
|---|---|---|---|
| Age | 1 | 76.4 | Continuous |
| Rac | 3 | Black | Categorical |
| Race - Hispanic | 1 | Hispanic | Categorical |
| Group Stage | 3 | III | Categorical |
| AJCC staging (T,N,M) | 13 | T1a | Categorical |
| Histology | 3 | Squamous cell neoplasms | Categorical |
| Metastatic sites | 4 | Metastasis to Bone | Categorical |
| Tumor Size | 1 | 1.5 cm | Continuous |
| Lymph Node Involvement[1] | 5 | Regional lymph nodes removed for examination with pre-surgical systemic treatment or radiation, but lymph node evaluation based on clinical evidence. | Categorical |
| AJCC Metastasis Eval [2] | 2 | Meets criteria for AJCC pathologic staging of distant metastasis | Categorical |

**Table 2**: Features used in the model, as well as aggregation used in data preprocessing before model training.

## 2.2. Machine Learning Model

### *2.1.1 Feature Construction*

Categorical features were one-hot encoded into separate features. For example, the feature race which includes categories (white, black, other) would be one hot encoded for a patient as (1,0,0) for white, (0,1,0) for black, and (0,0,1) for other. Continuous features were used in the current form. Standardization was performed on all features.

### *2.1.2 Predictive Modeling*

Principal component analysis was performed on the cohort to reduce dimensionality. Feature selection was performed using

[1] See NCI SEER definition: https://training.seer.cancer.gov/schema/rp_ureter/reg_ln_eval.html
[2] See NCI SEER definition: https://seer.cancer.gov/tools/ssm/2018-Summary-Stage-Manual.pdf

ANOVA F-value for the. A logistic regression23 model was trained using the features constructed, with the target labels

0: patient survived at least the 25th percentile of overall survival

1: patient death occurred before the 25th percentile of overall survival.

Grid search was performed to learn the most optimal set of modeling parameters from the following set: number of features {all}, regularization {l1, l2}, C {1e-2, 1e-1, 1, 1e1, 1e2}. The model was evaluated via 5-fold cross validation. The scikit-learn 24 Python package was used to implement the analysis.

# 3. RESULTS

## 3.1 Kaplan Meier Analysis

The median overall survival was 21 months. The 25th percentile of survival time was 11 months, which was used in the definition of patient classes for the machine learning problem (Figure 1).

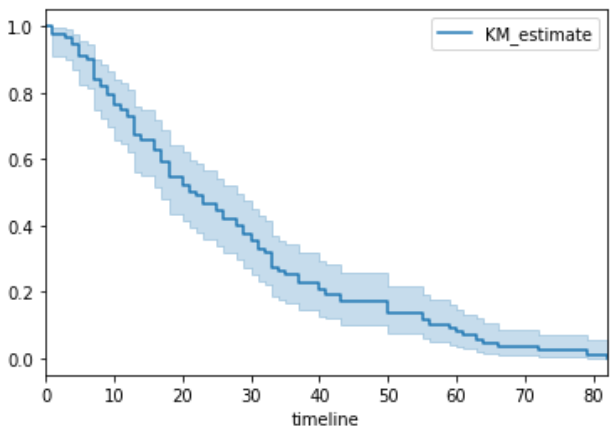



**Figure 1**: Kaplan Meier survival curve of all patients in the analytical cohort.

## 3.2 Principal Component Analysis

We utilized principal components analysis (PCA) to visualize the variation in the cohort of patients. Figure 2 shows the patients projected onto the first two principal components.

## 3.3 Predictive Model Performance Metrics

Across all folds of cross validation, the model achieved AUC of 0.696, accuracy of 0.682, precision of 0.381, recall of 0.405, and F1 score 0.375. The receiver operating curve is shown in figure 3.

## 3.4 Feature Importance

Of the 31 most predictive features with non-zero weights learned from the logistic regression model, all features conveyed clinical meaningfulness in the application of early mortality prediction. Higher age was correlated with early mortality, as well as patients whose histology are squamous cell neoplasms. On the other hand, patients' whose primary histology were nevi and melanomas, and transitional cell papillomas and carcinomas were correlated with lower early mortality. Interestingly, tumor size, hispanic race, american indian and asian race were negatively correlated with early mortality. Despite tumor size being negatively predictive of early mortality, T4 staging and group stage IV were the top and the third strongest positive predictive predictors. Features are visualized in terms of predicted weights in Figure 4.

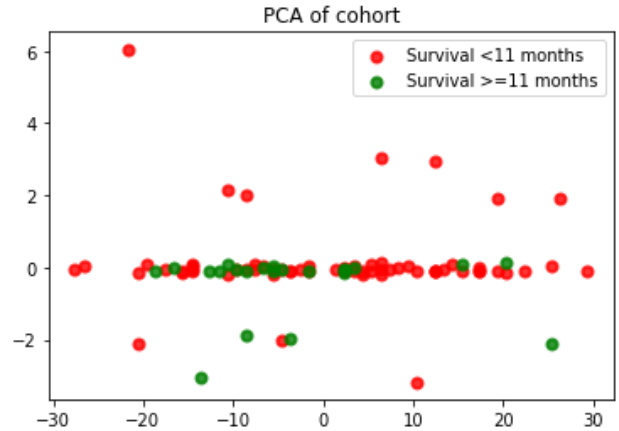



**Figure 2**: PCA plot of patients. Patients surviving less than 11 months are shown in red; otherwise green.

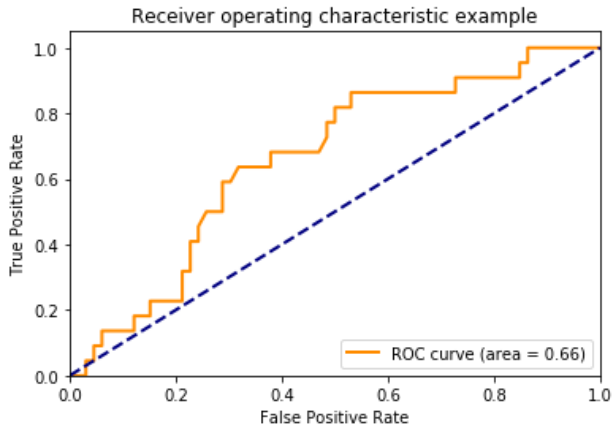


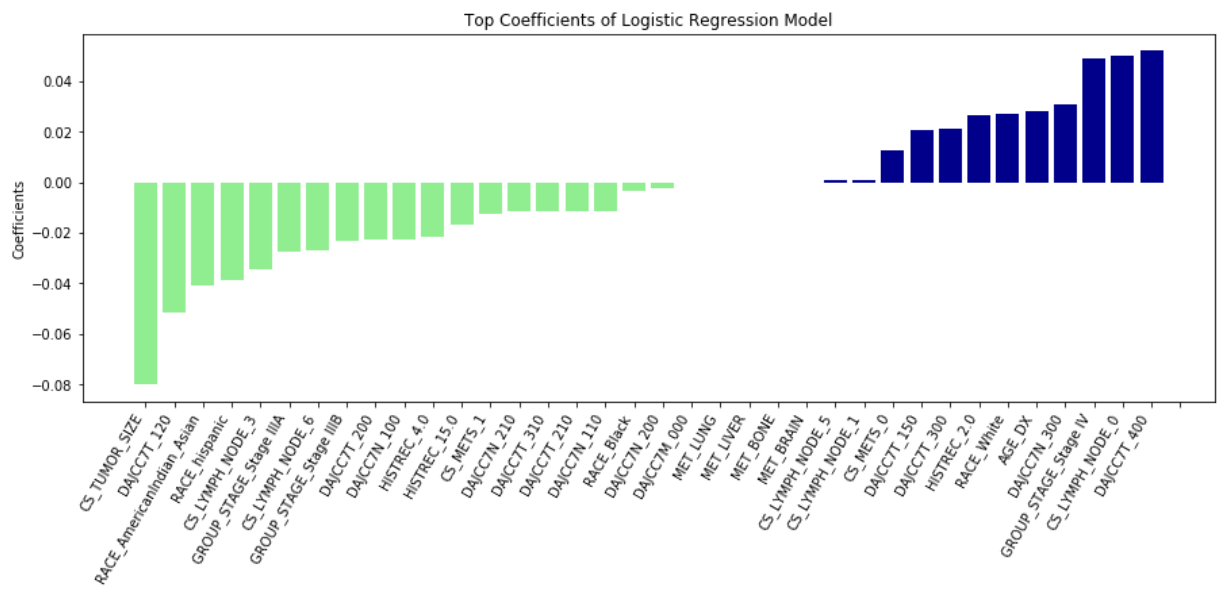

**FIgure 3**: Receiver operating curve for the logistic regression model.

# 4. DISCUSSION

A machine learning model was developed to predict early mortality in patients with stage III and IV penile cancer, using an end point of less than 11 months, equivalent to the median overall survival in advanced penile cancer patients.

The most predictive features were arrived at via extraction of weights learned from the logistic regression model. It is important to note that the most predictive features were determined from feature coefficients in the model. In our logistic regression model, feature importances were interpreted based on coefficients learned from the model. In clinical applications,

interpretability is important for downstream usage of the model in personalized treatment plans for patients. Methods such as




| Label | Feature Description | Weight |
|---|---|---|
| DAJCC7T_400 | Staging: T4 | 0.05198891 |
| CS_LYMPH_NODE_0 | No regional lymph nodes removed for exam | 0.04995687 |
| GROUP_STAGE_Stage IV | Group stage IV | 0.0486000377509148 |
| DAJCC7N_300 | Staging: N3 | 0.0306787043784349 |
| AGE_DX | Age at diagnosis | 0.02801539 |
| RACE_White | Race, White | 0.02719885 |
| HISTREC_2.0 | Histology: Squamous cell neoplasms | 0.02634187 |
| DAJCC7T_300 | Staging: T3 | 0.02083762 |
| DAJCC7T_150 | Staging: T1b | 0.02071698 |
| CS_METS_0 | Metastases: Evaluation of distant metastasis based on physical examination, imaging examination, and/or other non-invasive clinical evidence. No microscopic examination of metastatic specimen performed or microscopic examination was negative. | 0.01246206 |
| CS_LYMPH_NODE_1 | AJCC pathologic staging: No regional lymph nodes removed for examination. Evidence based on endoscopic examination, or other<br>invasive techniques including surgical observation without biopsy. No autopsy evidence used.<br>OR<br>Fine needle aspiration, incisional core needle biopsy, or excisional biopsy of regional lymph nodes or sentinel nodes as part of the diagnostic workup, WITHOUT removal of the primary site adequate for pathologic T classification (treatment). | 0.00089389 |
| CS_LYMPH_NODE_5 | AJCC pathologic staging: Regional lymph nodes removed for examination AFTER neoadjuvant therapy AND lymph node evaluation based on clinical evidence, unless the pathologic evidence at surgery (AFTER neoadjuvant treatment) is more | 0.00056434 |

| | | |
|---|---|---|
| | extensive. | |
| DAJCC7N_200 | Staging: N2 | -0.002198 |
| RACE_Black | Race, Black | -0.0036094 |
| DAJCC7T_210 | Staging: T2a | -0.0116877 |
| DAJCC7N_110 | Staging: N1a | -0.0116877 |
| DAJCC7T_310 | Staging: T3a | -0.0116877 |
| DAJCC7N_210 | Staging: N2a | -0.0116877 |
| CS_METS_1 | Metastases: No microscopic examination of metastatic specimen performed or microscopic examination was negative. | -0.0124621 |
| HISTREC_15.0 | Histology: nevi and melanomas | -0.0166282 |
| HISTREC_4.0 | Histology: transitional cell papillomas and carcinomas | -0.0217035 |
| DAJCC7N_100 | Staging: N1 | -0.0226636 |
| DAJCC7T_200 | Staging: T2 | -0.022684 |
| GROUP_STAGE_Stage IIIB | Group Stage IIIB | -0.0231021 |
| CS_LYMPH_NODE_6 | AJCC pathologic staging: Regional lymph nodes removed for examination AFTER neoadjuvant therapy AND lymph node evaluation based on pathologic evidence | -0.0271808 |
| GROUP_STAGE_Stage IIIA | Group Stage IIIA | -0.0275563 |
| CS_LYMPH_NODE_3 | AJCC pathologic staging: Any microscopic assessment of regional nodes (including FNA, incisional core needle bx, excisional bx, sentinel node bx or node resection), WITH removal of the primary site adequate for pathologic T classification (treatment) or biopsy assessment of the highest T category.<br>OR<br>Any microscopic assessment of a regional node in the highest N category, regardless of the T category information | -0.0346316 |
| RACE_hispanic | Race, Hispanic | -0.0389454 |
| RACE_AmericanIndian_Asian | Race, American Indian or Asian | -0.0409848 |
| DAJCC7T_120 | Staging: T1a | -0.0513743 |
| CS_TUMOR_SIZE | Tumor size | -0.0798414 |

**Figure 4**: top most predictive features, including all features with a positive weight or negative weight learned from the model. Positive weights indicate positive correlation with early mortality, while negative weights indicate negative correlation with early mortality.

LIME [25] have been successfully used in healthcare applications including prediction of mortality in ICU patients [15].

It is important to note that there are methods for identifying risk factors as correlated to mortality, such as Cox proportional hazards model. We did not implement a Cox model in this scenario because the problem was setup as a prediction problem.

It is interesting to note several interesting trends with respect to the predictive features identified. Higher age was correlated with early mortality, as well as patients whose histology are squamous cell neoplasms. On the other hand, patients' whose primary histology were nevi and melanomas, and transitional cell papillomas and carcinomas were correlated with lower early mortality. Interestingly, tumor size, hispanic race, american indian and asian race were negatively correlated with early mortality. However, despite smaller tumor size being predictive of early mortality, staging criteria indicative of more advanced disease were strongly predictive of early mortality (stage T4, group stage IV).

It is important to note that there are limitations of the study. First, there was a relatively small cohort size of 88 patients included. Many machine learning models for mortality prediction in other domains include significantly more patients. Despite this, the model was able to achieve AUC 0.65, which is consistent with performance of models for similarly challenging healthcare-related prediction problems [9,26].

We suggest 4 areas for future work. First, given the rare prevalence of penile cancer in the general population, we suggest thorough investigation of biomarkers associated with penile cancer. Biomarkers could include genomic, metabolomic, or microbiomic biomarkers. One example of a biomarker is

circulating tumor DNA (ctDNA), which has been implicated in prognosis of neoplasms such as non Hodgkins lymphomas including follicular lymphoma [27,28]. Second, we suggest studies assessing the effectiveness of the model in a prospective setting. Third, we suggest exploration from public tumor registrics such as the Cancer Genome Atlas [29], the Catalogue of Somatic Mutations in Cancer [30, 31], and the OncoKB Precision Oncology Knowledge Base [32]. Finally, we suggest unsupervised learning approaches to further characterize penile cancer. Approaches based on tensor factorization [33–38] may prove to be useful for uncovering distinct phenotypic subtypes of disease, as demonstrated in heterogeneous diseases such as heart failure.

# 5. CONCLUSION

We developed a machine learning model that predicts early mortality in penile cancer patients with 68.2% accuracy with AUC 0.696, and is able to identify clinical features predictive of early mortality. Future work should include integration of additional data sources, as well as explore temporal modeling strategies to account for clinical changes over time.

# 6. REFERENCES


[1] Barski, D., Georgas, E., Gerullis, H. & Ecke, T. Metastatic penile carcinoma – an update on the current diagnosis and treatment options. Central European Journal of Urology vol. 67 (2014).

[2] Compérat, E. Epidemiology and Histopathology: Penile Cancer. Urologic Oncology 1–8 (2019) doi:10.1007/978-3-319-42603-7_33-1.

[3] Bleeker, M. C. G. et al. Penile cancer: epidemiology, pathogenesis and prevention. World J. Urol. 27, 141–150 (2009).

[4] Goodman, M. T., Hernandez, B. Y. & Shvetsov, Y. B. Demographic and pathologic differences in the incidence of invasive penile cancer in the United States, 1995-2003. Cancer Epidemiol. Biomarkers Prev. 16, 1833–1839 (2007).

[5] Colón-López, V. et al. Penile cancer disparities in Puerto Rican men as compared to the United States population. Int. Braz J Urol 38, 728–738 (2012).

[6] Slopnick, E. A. et al. Racial Disparities Differ for African Americans and Hispanics in the Diagnosis and Treatment of Penile Cancer. Urology 96, 22–28 (2016).

[7] Ng, K., Steinhubl, S. R., deFilippi, C., Dey, S. & Stewart, W. F. Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time Before Diagnosis, Data Diversity, Data Quantity, and Data Density. Circ. Cardiovasc. Qual. Outcomes 9, 649–658 (2016).

[8] Chen, R., Stewart, W. F., Sun, J., Ng, K. & Yan, X. Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data: Implications for Temporal Modeling With Respect to Time Before Diagnosis, Data Density, Data Quantity, and Data Type. Circ. Cardiovasc. Qual. Outcomes 12, e005114 (2019).

[9] Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inform. Assoc. 24, 361–370 (2017).

[10] Rasmy, L. et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J. Biomed. Inform. 84, 11–16 (2018).

[11] Makino, M. et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci. Rep. 9, 11862 (2019).

[12] Johnson, A. E. W. et al. Machine Learning and Decision Support in Critical Care. Proc. IEEE Inst. Electr. Electron. Eng. 104, 444–466 (2016).

[13] Yu, C., Liu, J. & Nemati, S. Reinforcement Learning in Healthcare: A Survey. arXiv [cs.LG] (2019).

[14] Ahmad, M. A., Eckert, C. & Teredesai, A. Interpretable Machine Learning in Healthcare. in Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics 559–560 (Association for Computing Machinery, 2018).

[15] Katuwal, G. J. & Chen, R. Machine Learning Model Interpretability for Precision Medicine. arXiv [q-bio.QM] (2016).

[16] Chen, R. et al. explICU: A web-based visualization and predictive modeling toolkit for mortality in intensive care patients. Conf. Proc. IEEE Eng. Med. Biol. Soc. 2015, 6830–6833 (2015).

[17] Chen, R. et al. Cloud-based Predictive Modeling System and its Application to Asthma Readmission Prediction. AMIA Annu. Symp. Proc. 2015, 406–415 (2015).

[18] Desautels, T. et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. BMJ Open 7, e017199 (2017).

[19] Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 1, 18 (2018).

[20] Cheng, F. & Zhao, Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J. Am. Med. Inform. Assoc. 21, e278–e286 (2014).

[21] Hankey, B. F., Ries, L. A. & Edwards, B. K. The surveillance, epidemiology, and end results program: a national resource. Cancer Epidemiol. Biomarkers Prev. 8, 1117–1121 (1999).

[22] Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. J. Am. Stat. Assoc. 53, 457–481 (1958).

[23] Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. (Springer Science & Business Media, 2009).

[24] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).

[25] Ribeiro, M. T., Singh, S. & Guestrin, C. ‘ Why should i trust you?’ Explaining the predictions of any classifier. Proceedings of the 22nd ACM (2016).

[26] An, S. et al. Predicting drug-resistant epilepsy—a machine learning approach based on administrative claims data. Epilepsy Behav. 89, 118–125 (2018).

[27] Delfau-Larue, M.-H. et al. Total metabolic tumor volume, circulating tumor cells, cell-free DNA: distinct prognostic value in follicular lymphoma. Blood Adv 2, 807–816 (2018).

[28] Spina, V. et al. Circulating tumor DNA reveals genetics, clonal evolution, and residual disease in classical Hodgkin lymphoma. Blood 131, 2413–2425 (2018).

[29] Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013).

[30] 30. Forbes, S. A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 39, D945–50 (2011).

[31] Griffith, M. et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat. Genet. 49, 170–174 (2017).

[32] Chakravarty, D. et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol 2017, (2017).

[33] Wang, Y. et al. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. KDD 2015, 1265–1274 (2015).

[34] Ho, J. C., Ghosh, J. & Sun, J. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 115–124 (Association for Computing Machinery, 2014).

[35] Ho, J. C. et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. J. Biomed. Inform. 52, 199–211 (2014).

[36] Luo, Y., Wang, F. & Szolovits, P. Tensor factorization toward precision medicine. Brief. Bioinform. 18, 511–514 (2017).

[37] Perros, I. et al. SPARTan: Scalable PARAFAC2 for Large & Sparse Data. in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 375–384 (Association for Computing Machinery, 2017).

[38] Perros, I. et al. SUSTain: Scalable Unsupervised Scoring for Tensors and its Application to Phenotyping. in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2080–2089 (Association for Computing Machinery, 2018).