TLS report

Yves Greatti, Thomas Greatrix, Smurity Karale

June 2025

While Machine Learning (ML) techniques are used in the considered literature, the vast majority of papers use them only for statistical analysis. This, in turn, is done to identify predictors of TLS and TLS-related complications. The exceptions to this are , [1], [2], and [3]. Our analysis begins by exploring [1], [2], and [3] individually, and then explores how ML techniques have been used to perform statistical analysis to locate predictors of TLS.

1 ML for prediction of TLS

[1] Took data from 194 patients between the ages of 18 and 86 (of which 19 also had TLS) with Acute Myelogenous Leukemia (AML) and performed statistical analysis on this data in order to find predictors of Tumor Lysis Syndrome (TLS). The authors performed both univariate and multivariate analysis, which were done via logistic regression, a binary classification model, which works by fitting a logistic curve that splits the data into two user-defined categories, in this case, patients that have TLS and patients that do not have TLS. By looking at how important specific variables are to the model's predictions, one can infer the likelihood that a variable is a predictor of TLS. In this case, the authors found that uric acid "(UA) (p=.0003), Cr (p=.0025), [lactate dehydrogenase] LDH (p=.0001), [White Blood Cell count] WBC (p =.0058), gender (p =.0064), and [Chronic Myelomonocytic Leukemia] CMML (p=.0292)" were predictors of TLS. When the authors used multivariate analysis, LDH "(p=0.01, OR 3.01, 95% CI 1.5–6.2) and UA p =0.01, OR 2.00, 95% CI 1.4–2.8)" were predictors. While these values are statistically significant, the authors did not mention effect size.

There is some debate on how much data is required for logistic regression, a general rule of thumb is given by:

$$\frac{10C}{P}$$

Where C is the number of covariates (potential predictors of TLS in this case) and P is the smallest proportion of negative or positive cases. It is not stated precisely how many predictors were tested at once during multivariate analysis, but as 6 were found, we can assume $C \geq 6$. Hence:

$$P = \frac{19}{194} \approx 0.1$$
$$\frac{10C}{P} \approx \frac{10 \times 6}{0.1} = 600$$

The study had 194 datapoints and used less data than might be recommended for their multivariate logistic regression analysis. When using less data than recommended, there is a risk of the results not being representative of the general population; however, in some cases (especially within healthcare), such a situation is not always avoidable. Their univariate logistic regression did not have this issue, as for C = 1, we find that approximately 100 datapoints would be recommendable. As such, there may be precedence for UA, LDH, WBC, sex, and CMML to be predictors of TLS.

After using statistical analysis to find predictors of TLS, the authors used LDH and UA to develop a points-based predictive model of TLS (PPS-TLS score). Points were assigned based on LDH and UA values, with a bias given to UA. The PPS-TLS score varies between 0 and 6, and is challenging to

use it as a binary classifier: if you set a low threshold (e.g., score $\geq$ 2 or 3), you will catch all the TLS cases but you'll have so many positives that the test doesn't significantly change the pre-test odds (LR+ $\sim$ 1–1.7). The median score for patients with TLS was 5, with a sensitivity of 0.63 and a specificity of 0.85, indicating that the score has modest discriminative ability. A more rigorous approach would be to plot the whole ROC curve (and report the AUC), choose the threshold that maximizes Youden's J index, and then assess calibration or validate that cut-off in an independent sample.

Whilst it is easy to interpret a system wherein predictors are assigned points, the accuracy of the model and the fact that it was based on only two predictors, identified via multivariate analysis on a small dataset, mean that it is unlikely to be helpful in a clinical setting. The authors also state that the model has not been externally validated.

[2] took 772 adult patients, of whom 130 had TLS. TLS patients were divided into laboratory TLS (LTLS) or clinical TLS (CTLS). The authors began by categorizing all their continuous variables into categorical intervals before conducting a univariate analysis to identify important variables, which was accomplished using a chi-squared test. Afterwards, the statistically significant variables were used to build a stepwise logistic regression model. Using our rule of thumb from earlier:

$$P = \frac{130}{772} \approx 0.17$$

$$\frac{10C}{P} \approx \frac{10 \times 9}{0.17} \approx 529$$

Hence, the authors likely had enough data to create such a model. From their multivariate analysis, the authors found that WBC, UA, and LDH were predictors of TLS. The CTLS scoring system was established based on the regression coefficients of the multivariate analysis and has good performance (AUC. .81 %, 95% biascorrected CI, 0.77 to 0.84) on the test dataset, although goodness of fit Hosmer-Lemeshow statistic was not significant ($\tilde{\chi}^2$=7.6; p = 0.18).At cutoff levels at 2 and 3 points, the model achieved sensitivities of 95% and 89%, and specificities of 67% and 80%, respectively, for predicting clinical TLS. Although assigning discrete points to each predictor makes the score intuitive, its overall accuracy is not modest; it may be better deployed within a broader, multifactorial diagnostic framework.

[3] took data from 2,243 patients under 18 with acute lymphoblastic leukemia (ALL), of which 199 had TLS. The authors used this data to train a variety of machine learning models to predict whether a patient has TLS. Features with missing rates $>$ 30% were excluded (eliminating variables such as MBI). For those with missing rates $\geq$ 30%, the group used a miss-forest model to fill in missing values within the dataset. Data imbalance was addressed using an oversampling technique (SMOTE).

To start the study, the authors used LASSO regression to find predictors of TLS. LASSO regression is also known as L1 regression. LASSO regression minimizes the magnitude of coefficients in a Machine Learning model by penalizing the model, adding the absolute value of the coefficients multiplied by a coefficient to the model's loss. At a high level, this incentivizes the model to send the coefficients of unimportant variables to 0. As a result, by performing the LASSO regression, it is possible to tell which variables are insignificant as they will have been set to 0 during the LASSO regression. It is important to note that if two variables are highly correlated (e.g., age and weight in children), LASSO regression may shrink one to zero, even if both contribute value to the model. Additionally, selecting variables set to zero depends on the training data, making it challenging to interpret why LASSO prioritizes certain features over others. The authors of the paper minimized these disadvantages by using 10-fold cross-validation. However, there are still risks that should be considered when constructing models. The authors found that the LASSO regression selected FAB type, WBC, phosphorus, calcium, potassium, UA, AST, blood glucose, occurrence of infection, AKI, cardiac arrhythmia, and type of steroid used in the initial two induction chemotherapies as the 12 most important variables. WBC was also found to be a predictor of TLS in both studies (Xiao et al., 2024, [3] and Truong, 2007, [4]). In addition, Anthony R. Mato (2004) found that UA was a predictor of TLS, although this was in a study performed on adults rather than children, with a different form of cancer.

After using LASSO regression to find the 12 most important variables, the researchers created four different models to predict TLS from these variables: CatBoost, logistic regression, random forest, and a Support Vector Machine (SVM). Exploring multiple models is good practice, as it is often difficult to tell how a specific model structure will perform on a dataset without testing it. While it is commendable that they tested multiple hyperparameter configurations for each model to ensure fairness, the authors

do not appear to have evaluated the finalized models on various training datasets, ensuring no model was unfairly disadvantaged.

It is essential to recognize that model performance can be heavily influenced by the training data used and the initial conditions of a given run, which are typically random. Without repeated testing, there is a risk that model performance only varies due to data differences. This is especially relevant since the models analyzed by the authors performed similarly; therefore, changes to the training data and running repeated tests may have changed the results. It is also important to note that after identifying the twelve predictors of TLS, the authors did not investigate their biological and clinical significance. Without a strong biological argument, the paper loses some potential impact.

The best model was a CatBoost model with an AUC of 0.803 (95% CI : 0.735–0.865), indicating that TLS develops in all patients with initial white blood cell counts, high uric acid levels, and renal insufficiency, as demonstrated in previous studies. In addition, the authors found that the patients' glutamine aminotransferase, blood glucose, the occurrence of infection, acute kidney injury, cardiac arrhythmia, and the type of steroid used at the time of the initial induction chemotherapy were associated with the development of TLS . Shapley value analysis (SHAP) also revealed that among the 12 predictors used by CatBoost to predict TLS, low potassium and phosphorus are common, which is consistent with the state of the patients at initial diagnosis.

Ultimately, the paper demonstrates mostly good practices within AI and employs appropriate techniques. Although CatBoost outperformed the other models in the paper, this may not always hold true in practice. While CatBoost predictions are theoretically interpretable, understanding the reasoning behind its predictions in practice can be challenging, which may limit its practical applicability.

2 ML for statistical analysis of TLS factors

Within the considered literature, the vast majority of authors are focused on finding predictors of TLS and TLS-related complications. The methods used for this are typically well-established statistical techniques, with a heavy emphasis on logistic regression and, to a lesser extent, COX regression. Below, we have provided a table that provides information about each of the papers. Note that [1], [2] and [3] were discussed earlier and are thus excluded from the table.

Table 1: Summary of TLS Prediction Models (Part 1)

Name	Model(s) used	# Patients	# TLS Patients	Finding	Notes
Features at presentation... (Truong, 2007)	Logistic regression (univariate and multivariate)	74	328	Sex, age, WBC, mediastinal mass, hepatomegaly, splenomegaly, and T-cell immunophenotype were all predictors of TLS.	N/A
Serum phosphate level... (Lemerle, 2022)	COX regression	120	120	Increases in serum phosphate and LDH are early and reliable biomarkers of AKI in TLS.	Focuses on AKI, not TLS predictors.
Tumor Lysis Syndrome... (Rios-Olais, 2024)	Logistic regression (multivariate), COX regression	138	42	UA, LDH, and male sex predicted clinical TLS. LDH and WBC predicted TLS.	N/A
Uric Acid and TLS... (Ejaz, 2015)	Multinomial logistic regression	183	48 LTLS + 10 CTLS	WBC is a better predictor of TLS than others.	N/A
Predictors for Severe TLS (Wirth, 2012)	Logistic regression	1327	N/A	Multiple predictors including age, sex, cancer type, and serum markers.	Focus on severe TLS prediction.
Risk Factors in Childhood Leukemia (Prasertsan, 2024)	Logistic regression (univariate, multivariate)	252	51 with TLS (24 CTLS)	Age, BMI, WBC, LDH, GFR, AST among others identified.	N/A
TLS in CLL with Flavopiridol (Blum, 2011)	Logistic regression (univariate, multivariate)	116	53	WBC, gender, bulky lymphadenopathy, β2-microglobulin, albumin significant.	N/A
In-Hospital Outcomes Study (Durani, 2017)	Logistic + linear regression	28,370	28,370	Predictors included age, comorbidity score, insurance, hospital type, cancer type.	N/A

Table 2: Summary of TLS Prediction Models (Part 2)

Name	Model(s) used	# Patients	# TLS Patients	Finding	Notes
Predictors of in-Hospital Mortality	Logistic regression (univariate and multivariate)	997	997	Independent predictors were cardiac dysrhythmias and sepsis.	N/A
Children at Low Risk for TLS (Bahoush, 2015)	Logistic regression (univariate and multivariate)	160	41	CNS/renal involvement and T-cell immunophenotype were strong predictors.	N/A
Risk-Based TLS Management (Gopakumar, 2018)	Logistic regression (multivariate)	224	41	Early hydration, urine output, rasburicase improve TLS management in low-resource settings.	N/A
Plasma UA and Rasburicase (Canet, 2014)	Logistic regression	60	40	Higher baseline uric acid linked to AKI. Smaller UA response indicates subclinical AKI.	N/A

3 Discussion

The most common issue among the papers considered is the small datasets. Many of the papers considered lacked sufficient data to make reliable claims using logistic regression or more traditional statistical testing. As such, they risk producing results that cannot be reproduced. It is worth noting that this problem is widespread within the medical field due to the ethical and logistical issues with data collection.

Most papers use logistic regression, possibly with some pre-processing, to identify potential predictors of TLS. Such an approach is well-founded from a mathematical standpoint. However, many assumptions are made when using logistic regression. First, logistic regression is limited in its ability to capture complex relationships between variables, which may result in the omission of more intricate relationships. Logistic regression is also sensitive to outliers. Given the small sample sizes used and the complexity of real-world datasets, it is not unreasonable to expect outliers to be present in the data, nor is it unreasonable to assume that such outliers might be difficult to detect reliably. The usage of logistic regression for univariate and multivariate statistical testing, which is common in the papers considered, is not always a good approach (especially in the univariate case), as the models created are typically more complex to interpret and reproduce, and often are less robust and reliable for statistical testing than more traditional statistical tests. The decision to develop models capable of making predictions about TLS outcomes and then not evaluating them on their ability to do so is also questionable, especially when the primary purpose of these models is to make such predictions.

Some papers also explored the use of Cox regression to determine how variables may affect the time it takes for important events to occur. COX regression, similarly to logistic regression, is also sensitive to outliers. COX regression also assumes that the effect a variable has is constant over time. As a toy example, COX regression may be able to tell you that people who smoke are more likely to develop lung cancer. However, COX regression may also tell you that someone who has been smoking for 5 years is just as likely to develop cancer as someone who started yesterday, only if you include time effects like time since smoking started or smoking status over time. While Cox regression is well-established, authors typically do not acknowledge or account for these drawbacks in their analysis, which could render results unreliable and overlook potentially useful details.

In some of the papers considered, researchers also explored models that predicted the likelihood of TLS occurring based on provided variables. These approaches were typically point-based systems, wherein variables were assigned "points" based on their value. The points would then be added up, and based on the total, the risk of a specific outcome could be calculated. This approach is inherently well-suited for clinical environments, as it is extremely easy to interpret the result and see the most significant contributing factors to the result. Unfortunately, such models were generally not accurate enough to be clinically useful as a standalone tool, and it is recommended that the models be used as part of a larger diagnosis procedure instead. [3] compared a range of well-established, but far less interpretable, AI techniques and found that CatBoost had the best performance for predicting TLS. However, this approach was still unlikely to be accurate enough to be clinically useful as a standalone tool.

Although the methods used in the reviewed papers are, in principle, explainable, interpreting why a model identifies specific relationships can be challenging. It is also worth noting that analyses involving large numbers of variables are inherently difficult to interpret, regardless of the modeling approach. Nonetheless, predictions made by interpretable models can be experimentally validated after the analysis has been completed. In practice, such models are often complex, may uncover unexpected associations, and can occasionally show spurious correlations. Therefore, performing reality checks on the results, as well as analyzing potential follow-up studies, ensures that the findings from these models can be valuable for advancing TLS research.

It is worth noting that it is challenging to predict how well the models created will perform in real-world clinical settings. The accuracy of the models is often not verified, and the data is not publicly released, making it difficult to validate the results. The papers also do not create more than one model, or use multiple randomised test/train sets, which is standard within the Machine Learning field to test reliability. The small datasets used mean that models may not perform well for the general population, and they have significant risks of unexpected behavior in edge cases and rare cases.

4 Conclusion

In the future, the field may want to consider developing more sophisticated points-based TLS prediction models, which could be particularly well-suited to clinical settings due to their high interpretability for both specialists and non-specialists. If this turns out to be unfeasible, the usage of ensemble learning (the usage of multiple different models to make predictions) and transfer learning (taking a large, generalist model and training it to work well in your specific domain) have both been shown to perform well in the sort of low-data situations that appear to be common within the field.

References

[1] Pau Montesinos, Ignacio Lorenzo, Guillermo Martín, Jaime Sanz, Maria Luz Pérez-Sirvent, David Martínez, Guillermo Ortí, Lorenzo Algarra, Jesus Martínez, Federico Moscardó, et al. Tumor lysis syndrome in patients with acute myeloid leukemia: identification of risk factors and development of a predictive model. *Haematologica*, 93(1):67–74, 2008.

[2] Anthony R Mato, Brett E Riccio, Li Qin, Daniel Heitjan, Alison Loren, Martin Carroll, David Porter, Donald Tsai, Edward Stadtmauer, Sasha Perl, et al. A predictive model for tumor lysis syndrome in acute myelogenous leukemia. *Blood*, 104(11):883, 2004.

[3] Yao Xiao, Li Xiao, Yang Zhang, Ximing Xu, Xianmin Guan, Yuxia Guo, Yali Shen, XiaoYing Lei, Ying Dou, and Jie Yu. Prediction of tumor lysis syndrome in childhood acute lymphoblastic leukemia based on machine learning models: a retrospective study. *Frontiers in Oncology*, 14:1337295, 2024.

[4] T. H. Truong, J. Beyene, J. Hitzler, et al. Features at presentation predict children with acute lymphoblastic leukemia at low risk for tumor lysis syndrome. *Cancer*, 110(8):1832–1839, 2007. doi: 10.1002/cncr.22990.