PREDICTING THE NEED FOR MEDICAL CARE AFTER TOXIN EXPOSURE USING SHAP-INTERPRETABLE GRADIENT BOOSTING

A PREPRINT

Hugo Lerogeron
Hospices Civils de Lyon &
Biometrics and Evolutionary Biology Laboratory
University Claude Bernard Lyon 1
Lyon, France
lerogeron.hugo@gmail.com

Michael Chary
Department of Emergency Medicine
Weill Cornell
New York, New York, USA
mic9189@med.cornell.edu

Laurent Gueguen
Biometrics and Evolutionary Biology Laboratory
University Claude Bernard Lyon 1
Lyon, France
Laurent.Gueguen@univ-lyon1.fr

Kim An Nguyen
Hospices Civils de Lyon &
Biometrics and Evolutionary Biology Laboratory
University Claude Bernard Lyon 1
Lyon, France
huu-kim-an.nguyen@chu-lyon.fr

ABSTRACT

Objective: Experts in poison control centers must accurately and efficiently assess the severity of an exposure, neither delaying care nor pointlessly sending patients to the hospital, using only the information given during a first phone call. To help healthcare professionals (HP) make these difficult decisions, we developed and evaluated a machine learning-based algorithm that predicts whether a patient should seek medical help or not, based solely on the information provided during their first call to the poison control center, for all kinds of mono-intoxications.

Methods: We extracted data recorded by clinicians at the Lyon PCC between 2000 and 2025. Cases with missing original recommendations were excluded. We trained and compared several machine-learning models, emphasizing decision-tree–based and gradient-boosted tree approaches. Two classification tasks were defined: (1) binary triage (recommend emergency or non-emergency healthcare facility vs. stay at home) and (2) three-class triage (stay at home / non-emergency healthcare facility / emergency healthcare facility). Missing data were left as-is. Cross-validation and bootstrapping were used to ensure stable and statistically significant results. Model explainability was assessed with SHAP to identify the most important features for predictions. Model performance was evaluated using F1-score and ROC AUC; class imbalance was addressed during training. We compared our results to published algorithms that focus on single-substance intoxications.

Results: After processing, 220,825 cases remained. Recommended dispositions were: stay at home 66.6%, emergency facility 25.4%, and non-emergency facility 7.4%. For the binary task, XGBoost achieved the best performance (F1 = 0.748; ROC AUC = 0.820). For the three-class task, XGBoost again performed best (macro F1 = 0.657; multiclass ROC AUC = 0.859). The delay from exposure to call, SNOMED symptom codes, and the circumstance of exposure were the most influential features. Our results were competitive with algorithms focusing on intoxication due to a single substance.

Conclusion: Gradient-boosted tree models can produce accurate, interpretable, and clinically relevant predictions of poisoning severity from routine PCC data. With external validation and prospective testing, such tools could complement expert judgment to improve triage consistency and patient outcomes.

1 Background and objectives

In France, when someone has been exposed to a potentially toxic substance, they or their caregiver calls the nearest poison control center (PCC) to inquire about the steps necessary to deal with the possible intoxication. A healthcare professional (HP), using their knowledge and experience, must then assess the severity of the intoxication. To do that, they will ask the caller various questions, gathering data about the patient such as their demographics (e.g., age, weight, height), visible symptoms, type of substance and dose, or the time between exposure and the call. Despite possibly inaccurate or missing information, the HP must then recommend that the patient stay at the site of exposure to deal with the intoxication, seek medical assistance, or rush to get urgent medical care when necessary. However, making this decision is a challenging task. Every existing biological compound could be the cause of intoxication, and people react differently to compounds depending on the dose. In particular, in France, children represent the majority (55%) of cases, requiring specialized knowledge. Although cases are mostly mild (67.6%), 0.1% of cases result in death, making it difficult for HPs to correctly assess their gravity. The volume of cases is also significant: the Lyon PCC gets 80 to 100 calls every day. The first-line responders in PCC in France are the new HPs, like medical students, residents, pharmacists, pharmacy students, and nurses, who are not experts in toxicology. These human resources are renewed almost every 3 to 6 months. Therefore, the need to develop efficient tools to help new HPs in clinical toxicology is clear. In the era of large-scale health record databases collected in real-time and the increasing use of machine learning applications, we believe it is time to develop such tools. In particular, being able to classify whether a patient needs emergency treatment or not could help practitioners focus on the most difficult cases, while relieving pressure on emergency departments.

To help HPs make faster, more accurate decisions, various decision-support algorithms have been proposed. They are separated into categories: conventional approaches, rule-based reasoning, case-based reasoning, and machine learning-based reasoning.

1.1 Conventional Approaches

Conventional approaches refer to models that compute the severity of a pathology directly based on the clinical effects of the patient [1]. In the context of poisoning severity estimation, the most standard is the Poison Severity Score (PSS)[2], adopted by the IPCS, the Commission of the European Union, and the European Association of Poison Centers and Clinical Toxicologists (IPCS/EC/EAPCCT) [3]. The model assigns a grade to the case based on observed symptoms. It is not a prognostic score, but rather a tool to assess the severity level when clinical features are at their most critical: a higher grade will at least lead to a follow-up of the case by HPs[2]. The PSS is often used as a gold standard, either to ensure that newly developed scores are correlated to it [4], or used as a severity prediction model and compared with other scores [3], [5], [6]. Those studies show that the PSS is usually one of the best indicators of severity, by simply predicting that a PSS > 2 leads to a need for medical care in a hospital. On the other hand, it requires, by design, an accurate diagnosis done by an HP, which may not be possible through a phone call. The PSS is also used infrequently and misused, limiting its clinical utility[7].

Therefore, various other scores have been proposed in the literature, such as the Tanta University Risk Model (TURM). It is defined as follows :

$$\begin{aligned} TURM = &-1.966\ \textit{Glasgow Coma Scale (GCS)}\ - 0.329 * \textit{oxygen saturation} \\ &- 0.212 * \textit{diastolic blood pressure} + 0.27 * \textit{respiratory rate} - 0.33 + \textit{standard bicarbonate} \end{aligned} \quad (1)$$

It has been shown to have good performance (AUC = 0.79, N = 293) when predicting the need for ICU for acutely poisoned patients [8]. Flack and colleagues obtained similar results (AUC = 0.87, N = 400) using the same model [9].

Conventional approaches can also be used as a tool to understand the clinical effects that lead to a higher risk for patients. In [10], a multivariate logistic regression is used to highlight the main factors of poor outcomes for patients intoxicated by methanol, such as time from intoxication to treatment, respiratory arrest, and higher concentration of blood glucose. Brandenburg et al. [11] introduce a model that predicts the need for ICU in intoxicated patients while also showing the most important variables: the ICU Requirement Score (IRS). To achieve this, they create a point system, with patients "earning" points according to their characteristics. For instance, an age over 55 would grant 10 points, and cirrhosis would grant 7. By assessing the system's performance, they were able to show that respiratory insufficiency, age over 55, and a GCS below 6 were the most significant factors for ICU need. Conversely, alcohol poisoning was the variable associated with a lower need for ICU care. This score was validated in [12], who found that a negative IRS "virtually excludes the need for ICU care", but a positive IRS value was not enough to assess the patient's risk. Using the IRS on a French dataset, Gharbi et al. [5] showed that the negative predictive value (no need for ICU) of IRS < 6 was 95%. Using only age, oxygen saturation, and point-of-care lactate, [13] developed the poisoning Early Warning Score (pEWS), and obtained an area under the curve of 0.896. While the strength of

ICU predictive models like IRS or pEWS for hospital patients is that they can be used for the majority of hospitalised intoxication cases, other scores have been developed for specific intoxications, like [4] for acute poisoning with central nervous system xenobiotics, [14] for acute antipsychotic poisoning, or [15] for ethanol poisoning.

1.2 Case-based reasoning

The paradigm of case-based reasoning focuses on reasoning from experience. To find a solution for a given problem, we look for the most similar problem in a database, and then adapt the solution [16]. Here, cases are patients, stored in databases with additional information such as their demographics and results of biological or imaging tests. The difficulty lies in the algorithm used to find the most similar case. To that end, Barigou et al.[17] use "toxidromes", defined as a set of symptoms and signs that direct the clinical examination towards a particular class of toxic [18]. To efficiently find the most similar case, they first remove all cases in the database that do not contain the same "toxidrome" as the new case. Then, cases are scored based on the number of symptoms and attributes they have in common with the new case: the most similar one is used as "solution". The other difficulty lies in the method used to adapt the solution to the new case. In the approach proposed by [17], this step is left to the HP, as is often the case when dealing with healthcare applications [19].

1.3 Expert models

Expert models, also known as rule-based models, use a set of predefined rules and logical statements to make predictions. An example of such a rule could be: *If a patient's temperature is over 39°C, then they have a high fever*. These rules are made by human experts, using their experience and clinical guidance. The main advantage of expert models is their intuitive interpretability; they have therefore been used for a variety of healthcare-related tasks: Alonso et al. [20] predict the evolution of the end-stage renal disease in subjects affected by immunoglobulin A nephropathy, Haas et al. [21] assess the in-hospital mortality in ICUs and Sanz et al. [22] diagnose cardiovascular diseases.

The main limitations of expert models lie in the rule-making step. Indeed, experts need to account for every possible case to design the system, which can be very time-consuming when dealing with complex applications. That is the case for intoxication severity assessment: an intoxication can be due to virtually every substance that exists, resulting in hundreds of possible clinical effects. As a result, few attempts have been made to use expert models in this context. Still, Chary et al. [23] use a probabilistic logic network (PLN) to diagnose acute poisoning. A PLN consists of pairs of probabilities and logical statements, acting as rules.

$$0.4\ \textit{somnolent} \longrightarrow \textit{sedative hypnotic}(x) \tag{2}$$

Equation 2 is an example of such pairs, showing that in 40% of cases, if a patient is somnolent, then they may have been intoxicated by an agent of the sedative hypnotic class. 34 of those pairs were made to predict the toxidrome of a generated case, showing performance similar to humans on easy cases, but underperformed compared to clinicians on more complex cases.

1.4 Machine Learning based reasoning

The goal of machine learning (ML) methods is to learn a model that automatically maps inputs to the desired outputs. Here, it would mean feeding the model patients' data to directly predict the severity of the intoxication. While conventional approaches and rule-based reasoning require expert knowledge for model development, ML models are complex algorithms designed without specific dependence on the actual use case. However, the model learns by trying to mimic the prediction of actual experts: their help is therefore still needed. This paradigm is called supervised learning [24]. ML approaches can leverage huge databases to learn their task, and can afterwards be used to predict for new patients. ML has been shown to outperform conventional approaches on ICU prognosis [1] and has recently been used on numerous toxin exposure predictions: [25] on acute methanol poisoning, [26] on acute organophosphate poisoning, [27] on diphenhydramine exposure, or [28] on acute diquat poisoning. In particular, models based on decision trees perform the best. The goal of decision trees is to learn implicit decision rules (logical rules of the form *if then*, generally comparing a variable to a threshold) from the data. By themselves, decision trees are easy to understand. In most applications, however, models are made of numerous decision trees, with each decision tree only considering a subset of the dataset, like in Random Forest. Decision trees are used as a set of weak learners, whose predictions are averaged to get a stronger, less biased prediction. When trees are trained iteratively to improve on the previous tree's mistakes, we get gradient boosting algorithms, such as XGBoost. These advanced algorithms trade higher performance for worse interpretability similarly to deep learning models.

Deep Learning (DL)[29], a subset of machine learning, uses artificial neural networks to learn complex patterns in the data. There exist various types of standard architectures for DL models, such as convolutional neural networks,

recurrent neural networks, dense neural networks, etc [30]. In the context of intoxication severity assessment, data is mostly tabular, which is still an area where deep learning models struggle to outperform algorithms based on decision trees [31]. While predicting the need for intubation for patients poisoned by methanol, [32] confirms that decision trees outperform various architectures of deep learning models.

Despite their performances, the use of ML models is limited by the difficulty of understanding their predictions. Indeed, conventional approaches are based on simple models; it is easy for an HP to understand why a prognosis prediction was made. In the same way, rule-based reasoning aims at mimicking the way HPs make a decision, and case-based reasoning justifies its prediction by the most similar known case. The explainability of those approaches is called *Ante-hoc*: HPs can directly understand the prediction [33]. On the other hand, ML algorithms are made of numerous complex mathematical operations, making their functioning akin to a 'black box'. As a result, HPs tend to have low trust in ML-based approaches and use the system less [34]. Therefore, ML methods require *Post-hoc* explanations techniques [33]. In the context of poison prognosis, explanation is mostly based on showing the importance of features on decisions [25], [32], [28], using tools like SHAP. SHAP (SHapley Additive exPlanations) is a method that assigns each feature a numerical value representing its contribution—positive or negative—to a model's prediction, based on principles from cooperative game theory[1].

Overall, while these approaches show various ways to help HPs improve poison prognosis, they all have the same limitations, namely that they require numerous biological data and are only used to predict the prognosis of a few toxins. To the best of our knowledge, no ML model currently exists that predicts poisoning severity using only information obtained during the initial phone call for every possible agent. Accordingly, this study presents an ML-based approach to predict the need for medical care, while simultaneously performing triage of minor cases in which exposed individuals can safely remain at home, using only data available during the first contact with a PCC. Our objective is to develop an algorithm capable of assessing intoxication severity across all substances represented in our dataset. We systematically evaluate multiple machine learning algorithms using comprehensive performance metrics to identify the optimal predictive approach. Additionally, we benchmark our method against existing approaches that assess intoxication severity for specific subsets of toxic substances, demonstrating the broader applicability and effectiveness of our proposed solution.

2 Materials and methods

The goal of our model was to fulfill the following objectives :

- Accurately predict whether a patient needs medical care or not, using only the information provided during the first call to the poison center.
- Provide as much information as possible to help the practitioner understand the prediction.
- Ultimately, reduce triage error.

2.1 Dataset

During each call, HPs systematically recorded comprehensive patient information, including demographics, presenting symptoms, medical history, circumstances of exposure, timing and location of the incident, and when available, the specific substance involved and its dosage. Based on this information, HPs provided recommendations to callers regarding the appropriate level of medical care required, ranging from no treatment necessary to emergency department referral. This recommendation served as our target variable for prediction. Cases lacking this critical outcome information were excluded from the analysis.

2.1.1 Data preprocessing

We developed a comprehensive data preprocessing pipeline to transform raw clinical data into a format suitable for machine learning analysis. The pipeline directly addressed the complex, multi-structured nature of medical records, which contain nested symptom details, agent information, and patient characteristics.

First, we filtered event records by extracting the event nature identifier from structured event details. We excluded cases classified as "AVIS" (advisory notifications) to focus the analysis on confirmed intoxicated events. This filtering step ensured that the final dataset included only clinically relevant cases. We also removed multiple entries for the same patient, keeping only the oldest record corresponding to the first call to the PCC. This approach simulated an operator receiving a call about a brand-new patient, as typically occurs in PCCs.

[1] https://shap.readthedocs.io/en/latest/

We then performed initial data cleaning by standardizing text fields and handling missing values. We corrected age calculations by replacing empty strings and "NaN" entries with proper null values, converting comma-separated decimals to dot notation, and removing physiologically unlikely ages (those over 125 or below 0 years). We applied similar validation to anthropometric measurements, removing weight values outside the 0–200 kg range and height values outside the 0–250 cm range. We encoded the symptom information using the Systemized Nomenclature of Medicine – Clinical Terms (SNOMED[35]) terminology. Because patients could present multiple symptoms simultaneously, we applied a MultiLabelBinarizer[2] to create binary symptom indicators. To address the high dimensionality of the resulting feature space (which contained more than a thousand features), we used Truncated Singular Value Decomposition (TruncatedSVD) to reduce the symptom representation to N_{dim} principal components while preserving the most significant symptom patterns. We determined N_{dim} through a grid search hyperparameter exploration with the following candidate values: $N_{dim} \in [1, 25, 50, 100]$.

We identified the primary causative agent through a hierarchical extraction process applied to the agent detail strings. The algorithm prioritized the most specific agent classification (level 7) and moved progressively to broader categories (levels 6–1) when more detailed information was unavailable. We implemented a special case for carbon monoxide exposures, which were frequently under-coded in the original data. We removed multi-intoxications (12.5% of the dataset) because this study focuses exclusively on mono-intoxications. We extracted exposure information from multi-level nested structures that contained the estimated amount, confidence in the estimation, exposure circumstances, exposure location, exposure type, exposure route, and exposure duration.

Finally, we converted all temporal measurements into a standardized minute-based format. This standardization enabled consistent temporal analysis across the varied measurement units present in the original data. We standardized healthcare facility information using a two-tier or three-tier classification system. The primary categories included home/exposure site (H), emergency facilities (EF), and non-emergency healthcare facilities (NEHF). This information served as the label we aimed to predict. For binary classification tasks, we merged EF and NEHF into a single category (HF).

2.2 Machine learning models

Following [36] and [25], we focused our efforts on decision tree-based algorithms: Random Forest (RF), XGBoost, LightGBM, and HistGradientBoostingClassifier(HBC)[3]. All the models had their default hyperparameters and were implemented using scikit-learn [37]. To fit and evaluate the algorithms, we used a 5-fold cross-validation method. Each algorithm was fitted on the train set and evaluated on the validation set for each fold. We used the following metrics to compare the models :

- Area Under the Receiver Operating Characteristic Curve (ROC AUC)[4]
- Precision : $\frac{TP}{TP+FP}$
- Recall (or sensitivity) : $\frac{TP}{TP+FN}$
- F1 score : $\frac{2*TP}{2*TP+FP+FN}$

We used the macro F1 score to account for the substantial data imbalance in our dataset, as it equally values precision and recall. This choice reflects our clinical priorities: underestimating the risk of intoxication (false negatives) could endanger patients, while recommending medical care for too many patients (false positives) could overwhelm healthcare facilities and increase waiting times, a factor shown to raise mortality [38]. Several approaches have been proposed to mitigate the effects of imbalanced data on predictive performance [39]. In our study, we evaluated four such methods: undersampling, training with balanced class weights, and oversampling using SMOTE [40] and ADASYN [41]. The sampling methods are only applied to the training sets.

2.3 Explanations of predictions

As mentioned earlier, being able to explain the predictions made by the algorithm will significantly enhance its usability for practitioners. Since machine learning models, and particularly gradient boosting algorithms, are complex, we used *Post-hoc* interpretation techniques. In detail, we used the importance the algorithms give to each variable to try

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html

[3] For RF and HBC, we used their scikit-learn implementation.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

to explain how the predictions were made. To complement this analysis, we used the SHAP library. In particular, we use the TreeSHAP algorithm[5] and display the mean of the absolute value of SHAP.

2.4 Error Analysis

To better understand the model's limitations, we conducted an error analysis by examining cases where the model's predictions deviated from ground truth labels. We isolated these misclassified instances and performed distributional comparisons for each input variable, contrasting their distributions in the error subset against those in the complete dataset. Categorical variables were analyzed using chi-square tests, while continuous variables were compared using two-sample Kolmogorov-Smirnov tests. We applied the Bonferroni correction for multiple comparisons. This approach allowed us to identify systematic patterns in the model's failure modes and potential biases in its decision-making process. To further characterize these systematic differences, we compared the marginal distributions of each variable between the error and full datasets. For each feature, we calculated the difference in relative frequency between the two sets and identified the category with the largest positive discrepancy (i.e., the category most over-represented among error cases).

2.5 Comparison with other approaches

Comparing results across different models presents significant challenges. Datasets vary between studies and are often unpublished; severity definitions differ across contributions, scoring methodologies are inconsistent, and most studies focus on restricted subgroups of toxins rather than comprehensive toxin panels. Therefore, to enable meaningful comparison with other contributions, we restricted our model's results to a specific set of agents by computing metrics exclusively on this defined subset. We also aligned the number of classification classes, comparing binary classification contributions using our binary approach and three-class contributions using our three-class approach.

3 Results

Feature Type	Feature Name	Description	Missing Values %	Distribution
Demographics	Age	As reported by the caller (years)	4.51	9.0 (0.0 - 2.0 - 36.5 - 125.0)
	Weight	As reported by the caller (kg)	82.63	16.0 (0.0 - 12.0 - 55.0 - 200.0)
	Height	As reported by the caller (cm)	99.50	163.0 (0.0 - 149.25 - 172.0 - 196.0)
	Sex	As reported by the caller	0.23	F(50.2%) - M(48.8%) - U(1.0%)
Exposure	Agent	As reported by the caller	1.65	Detergent(4.9%)
	Estimated amount	Quantitative estimation given by practician	35.62	Low(56.7%)
	Confidence in amount	Evaluation of the estimation	46.08	High(90.6%)
	Circumstance of exposure	Accidental, voluntary or unknown	0	Lack of risk perception(35.3%)
	Place of exposure	e.g. forest, garden, kitchen, ...	0	U(52.7%)
	Type of exposure	Acute, chronic, ...	0	Acute (98.2%)
	Route of exposure	e.g. ocular, respiratory, dermal, ...	0.7	Oral(68.3%)
	Duration of exposure	As reported by the caller (min)	53.90	Unknown (64.3%)
	Delay from exposure to call	As reported by the caller (min)	58.34	20.0 (0.0 - 10.0 - 97.0 - 86,400)
Symptoms	SNOMED $\times$ 25	25 dimensions of encoded SNOMED codes	0	N/A
	HasSymptoms	Boolean parameter	0	No(55.4%)
Label		Preconised place of care	0	Home(66.6%) - EHF(25.39%) - NEHF(7.4%)

Table 1: Features used by our models to predict the severity of intoxication. For numerical variables, distribution indicates the median (min - 25% - 75% - max) values. For categorical variables, we include the most frequent category. U means Unknown.

We conducted all experiments using Python 3.11. We identified the optimal N_{dim} as $N_{dim} = 25$. We kept 220,825 unique cases for our analysis. The preprocessing pipeline transformed the original medical records into a normalized tabular format suitable for machine learning applications, producing the data matrix $X \in \mathbb{R}^{220,825 \times 40}$. Table 1 lists every feature used by our models, along with its percentage of missing values. Because the RF, XGBoost, LightGBM, and HBC classifiers can natively handle missing values, we kept those values unchanged. We also highlight the severe data imbalance: clinicians suggested NEHF for only 7.4% of patients.

3.1 Binary classification experiment

The results of the binary classification using balanced weights are presented in Table 2. Overall, the gradient-boosted models (XGBoost, LightGBM, HBC) achieved comparable performance, with XGBoost performing slightly better.

[5] https://shap.readthedocs.io/en/latest/generated/shap.TreeExplainer.html

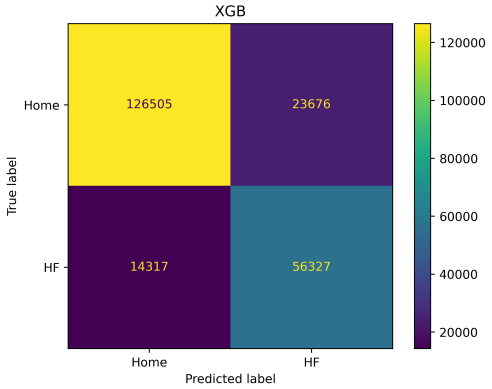


Figure 1: XGBoost's confusion matrix for the binary classification task.

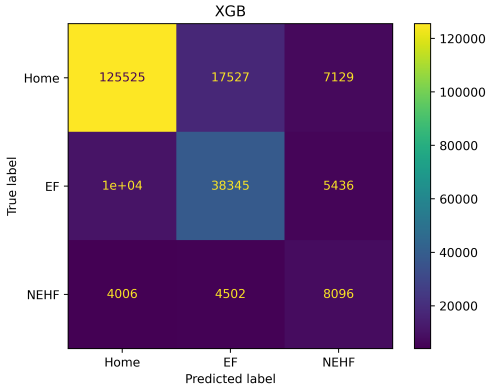


Figure 2: XGBoost's confusion matrix for the 3-class task.

An accuracy of 0.828 indicates a decent predictive capability and suggests potential for clinical application. Furthermore, all models demonstrated near-instantaneous inference times, supporting their feasibility for real-time use. The confusion matrix is shown in Figure 1.

Model	Accuracy	Precision	Recall	F1	RocAUC
RF	0.821 ± 0.001	0.802 ± 0.003	0.584 ± 0.005	0.676 ± 0.004	0.758 ± 0.002
XGBoost	0.828 ± 0.001	0.704 ± 0.001	0.797 ± 0.003	0.748 ± 0.001	0.820 ± 0.001
LightGBM	0.825 ± 0.001	0.694 ± 0.002	0.808 ± 0.002	0.747 ± 0.002	0.820 ± 0.001
HBC	0.821 ± 0.001	0.687 ± 0.004	0.812 ± 0.003	0.744 ± 0.001	0.819 ± 0.001

Table 2: Results of the classification distinguishing between patients needing medical care or not.

3.2 Three-class experiment

Results of the three-class experiment are given in Table 3. Once again, gradient boosting methods obtained similar performances, with XGBoost being slightly better. The higher dataset imbalance justified the lower F1-score. The confusion matrix is shown in Figure 2.

Model	Accuracy	Precision	Recall	F1	RocAUC
RF	0.788 ± 0.002	0.697 ± 0.005	0.563 ± 0.005	0.600 ± 0.005	0.840 ± 0.001
XGBoost	0.779 ± 0.001	0.642 ± 0.001	0.678 ± 0.001	0.657 ± 0.001	0.859 ± 0.001
LightGBM	0.755 ± 0.002	0.621 ± 0.001	0.691 ± 0.001	0.642 ± 0.002	0.859 ± 0.002
HBC	0.806 ± 0.001	0.642 ± 0.001	0.617 ± 0.001	0.647 ± 0.002	0.856 ± 0.001

Table 3: Results of the classification distinguishing between patients needing emergency care, medical care, or no care at all.

3.3 Sampling strategies

Table 4 summarizes the performance of different sampling strategies. Since XGBoost achieved the best results in the previous experiment, it was used for this comparison. For both tasks, the balanced weight strategy yielded the best results. All the scores were similar for the binary classification task, which was expected given the relatively mild class imbalance. In contrast, the three-class setting exhibited a stronger imbalance, where undersampling produced inferior results compared to the other methods.

Experiment	Undersampling	Balanced weights	SMOTE	ADASYN
2-class	0.744 ± 0.001	0.748 ± 0.001	0.736 ± 0.003	0.737 ± 0.002
3-class	0.614 ± 0.001	0.657 ± 0.001	0.651 ± 0.002	0.647 ± 0.003

Table 4: Comparison of F1-scores for different sampling strategies using XGBoost.

Table 5: Importance of feature for the classification using the built-in weights. The top 10 features are shown.

(a) Binary task

Feature	Importance
SNOMED	0.3235
Agent	0.1576
Age	0.1531
Delay from exposure to call	0.0867
Circumstance of exposure	0.0593
Estimated amount	0.0507
Weight (kg)	0.0466
Route of exposure	0.0347
Duration of exposure	0.0258
Sex	0.0236

(b) Three-class task

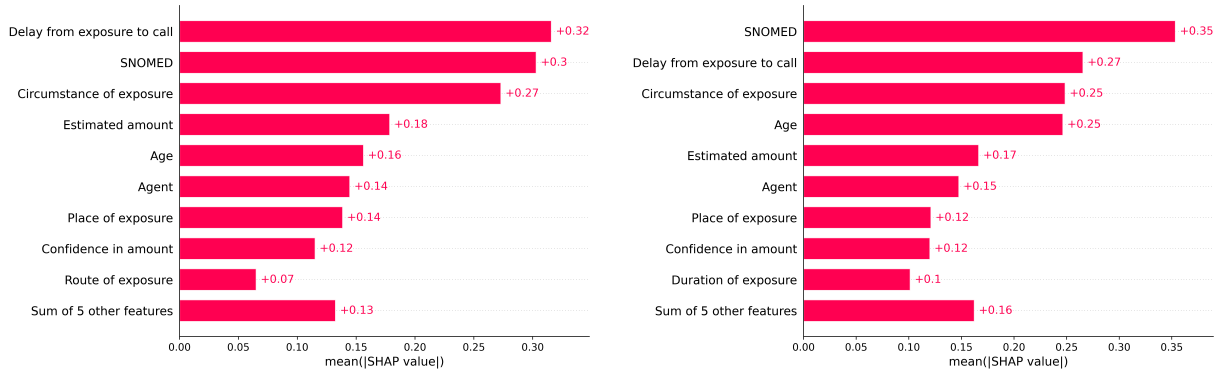
Feature	Importance
SNOMED	0.3326
Age	0.1837
Agent	0.1705
Delay from exposure to call	0.0689
Weight (kg)	0.0473
Circumstance of exposure	0.0434
Estimated amount	0.0403
Route of exposure	0.0289
Sex	0.0275
Duration of exposure	0.0191

3.4 Explanations of predictions

We used Xgboost's built-in importance weights to highlight the most important features, as shown in Table 5. In both tasks, the most important features were the SNOMED codes, the agent responsible for the exposure, and the age of the patient. We display the results obtained using SHAP, for both tasks, in Figure 3. This time, the SNOMED codes, delay from exposure to call, and circumstance of exposure are found to be the three most impactful features. Still, age and agent are found in the top six, and are therefore still useful to correctly predict the outcome.

3.5 Error Analysis

Results of the statistical tests are reported in the supplementary materials (Table 8). Nearly all variables exhibited significantly different distributions between the full dataset and the error subset. This indicates that the XGBoost model did not make errors at random; rather, certain types of cases were systematically harder to classify. Marginal distributions of each variable between the error and full datasets in Table 6. Overall, the model performed worse on cases with missing or insignificant information. Ocular exposures were more frequently misclassified in the binary task, which may reflect a tendency for healthcare professionals to be particularly cautious in such situations, whereas respiratory exposure was the most common route of exposure in errors for the three-class task.



(a) Binary task.

(b) Three-class task.

Figure 3: Variable importance using SHAP.

Table 6: Largest positive differences in distribution between the error subset and the full dataset. For each variable, the most over-represented category among error cases is shown.

(a) Binary task

Feature	Value	Difference
Place of exposure	Not significant	0.1378
Circumstance of exposure	Everyday accident	0.0795
Estimated amount	Unknown	0.0648
Duration of exposure	Unknown	0.0269
Route of exposure	Ocular	0.0229
Sex	M	0.0190
Confidence in amount	Medium	0.0067
Type of exposure	Sub-acute	0.0033

(b) Three-class task

Feature	Value	Difference
Place of exposure	Not significant	0.1793
Estimated amount	Unknown	0.1008
Circumstance of exposure	Everyday accident	0.0637
Route of exposure	Respiratory	0.0393
Sex	M	0.0270
Duration of exposure	Unknown	0.0258
Type of exposure	Sub-acute	0.0069
Confidence in amount	Weak	0.0052

3.6 Comparison with other approaches

The results are summarized in Table 7, and detailed in the following paragraphs.

3.6.1 Two class approach

Hosseini et al. [26] developed a binary classification model to predict organophosphate intoxication severity, distinguishing between mild-to-moderate and severe cases in a cohort of 1,237 patients. Their optimal model employed XGBoost and achieved an F1-score of 0.91, with the most predictive features being venous blood gas pH (VBG-pH), white blood cell count, plasma cholinesterase activity, venous blood gas base excess (VBG-BE), and patient age. Given the constraints of our dataset, we could not restrict analysis to specific chemical compounds; therefore, we included all toxin categories potentially containing organophosphates: "insecticide," "herbicide," "vegetal pesticide," and "animal pesticide." Additionally, our dataset lacked access to the biological parameters that proved most valuable in their study. Our analysis encompassed 2,490 cases, of which 1,028 (41.2%) required urgent care, yielding an F1-score of 0.76. The five most important predictive features in our model were the SNOMED codes, the circumstance of exposure, the delay from exposure to call, the estimated amount of dose, and the place of exposure.

Patel et al. [42] developed a predictive model for pediatric intensive care unit (PICU) admission following acute intoxication in a cohort of 70,364 patients stratified by age. Their dataset included 21,646 patients under 6 years (14.1% requiring PICU), 4,085 patients aged 6-12 years (16% requiring PICU), and 44,683 patients over 13 years (16.7% requiring PICU). The model achieved validation accuracies of 0.69, 0.53, and 0.57 for these respective age groups. For comparison, we filtered our dataset to match these age categories and evaluated our model's performance on predicting urgent care needs. Our cohort comprised 97,235 patients under 6 years (20.1% requiring urgent care), 11,410 patients aged 6-12 years (26.0% requiring urgent care), and 12,796 patients aged 12-18 years (37.8% requiring urgent care). Our model demonstrated superior performance with validation accuracies of 0.86, 0.85, and 0.84 across all three age groups, respectively.

Rose et al. [43] developed a predictive model for mortality following acute carbon monoxide intoxication in a cohort of 1,273 cases, of which 57 (4.5%) resulted in death. Using a clinical scoring system based on 10 predictors (excluding biological parameters), they achieved an AUC of 0.70 on the validation set. The most important features for their model were the age of the patient, the presence of cardiac complications, and an abnormal mental status. For comparison, we restricted our dataset to carbon monoxide cases, encompassing 9,986 cases with 53.0% requiring urgent care. Our model achieved a superior AUC of 0.82. The five most important predictive features in our analysis were the circumstance of exposure, SNOMED codes, delay from exposure to call, estimated amount, and place of exposure.

3.6.2 Three class approach

Mehrpour et al. [27] predicted the severity of diphenhydramine overdose using three classes: minor, moderate, and major. They constructed their dataset based on the National Poison Data System database for a total of 53,671 patients. They obtained an F1-score of 0.75, but did not report on the importance of features. For the same reasons as above, we restricted our results to "ATC R06" toxins. We had 2,135 cases. We obtained an F1-score of 0.62. The 5 most important features were the SNOMED codes, the circumstance of exposure, delay from exposure to call, age, and estimated amount.

Mehrpour et al. [44] predicted the severity of acetaminophen overdose using three classes: minor, moderate, and major. They constructed their dataset based on the National Poison Data System database for a total of 31,551 patients. They

obtained an F1-score of 0.76. The most important features were elevated serum levels of enzymes, other liver function test abnormalities, and anorexia. We restricted our dataset to "ATC N02" toxins and had 7,844 cases. We obtained an F1-score of 0.66. The most important features were the circumstance of exposure, SNOMED codes, delay from exposure to call, age, and estimated amount.

Table 7: Comparison of toxicology prediction models with current study results

Study	Toxin/Target	Cases	Classes	Metric	Performance	Our Cases	Our Performance
Two-Class Approach							
Hosseini et al. [26]	Organophosphate (mild-moderate vs severe)	1,237	2	F1-score	0.91	2,490 (41.2% urgent)	0.76
Patel et al. [42]	PICU admission	21,646 ($\leq 6y$) 4,085 (6-12y) 44,683 ($\geq 12y$)	2	Accuracy	0.69 0.53 0.57	97,555 ($\leq 6y$) 11,554 (6-12y) 12,798 (12-18y)	0.86 0.85 0.84
Rose et al. [43]	Carbon monoxide (mortality prediction)	1,273 (4.5% deaths)	2	AUC	0.70	9,986 (52.9% urgent)	0.82
Three-Class Approach							
Mehrpour et al. [27]	Diphenhydramine (minor/moderate/major)	53,671	3	F1-score	0.75	2,134	0.62
Mehrpour et al. [44]	Paracetamol (minor/moderate/major)	31,551	3	F1-score	0.76	7,844	0.66

4 Discussion

Our findings demonstrate that machine learning approaches, particularly XGBoost, can accurately predict the need for emergency care after intoxications at the initial contact with the poison control center. The best performance was obtained in the binary classification experiment, which aligns with the clinical reality that distinguishing clearly benign cases from those requiring urgent evaluation is often the most critical decision. The three-class setting proved more challenging, but results still show that the model retains useful discriminative capacity.

Feature importance analysis through built-in weights and SHAP confirmed that the model's decisions are clinically plausible. Symptoms at presentation, the circumstances of exposure, and the delay between exposure and call emerged as the strongest predictors. These findings are consistent with expert reasoning, providing reassurance about model validity and enhancing interpretability for healthcare professionals. Indeed, the high weight assigned to presenting symptoms reflects the "toxidrome" approach in emergency medicine. For instance, the presence of neurological depression or respiratory distress naturally shifts the model toward a high-risk classification, which is the gold standard in bedside triage. Concerning agents, the model correctly identifies that certain substances () have a lower threshold for emergency intervention regardless of initial symptoms, due to their known potential for rapid clinical deterioration. Finally, the significance of patient age validates the model's sensitivity to vulnerable populations. Pediatric accidental ingestions and geriatric polypharmacy cases present distinct risk profiles that the model successfully captured.

Compared with previous approaches relying on rule-based expert systems or conventional scoring methods such as the PSS, our model offers three main advantages: higher predictive performance across a large and heterogeneous dataset, the ability to adapt to diverse poisoning scenarios without manual rule-encoding, and transparent explanations via SHAP that may foster user trust. This balance between performance and interpretability addresses a key barrier to the adoption of machine learning in clinical toxicology.

Nevertheless, some limitations should be acknowledged. First, the dataset was restricted to single-substance exposures, whereas multi-intoxications remain common in practice and often drive the most severe cases. Second, the data come from a single centre in Lyon, which may limit generalisability to other PCCs with different populations or recording practices. Future work should therefore include multi-centre datasets and integration of poly-intoxications. Third, our algorithm obtained worse results on the non-emergency healthcare facilities class. While this could be explained by the data imbalance, it could also stem from the fact that those cases are inherently difficult to classify, even for practitioners. Biomarkers may be necessary to classify those gray-area cases better.

From a clinical implementation perspective, the proposed model could be integrated as a real-time decision-support tool within poison control centers, assisting specialists in rapidly stratifying risk and prioritizing cases requiring urgent referral. By providing an additional layer of objective, data-driven assessment, it may help standardize triage across different operators, reduce variability in advice, and ensure that high-risk patients are identified without unnecessary delay. In emergency departments, such models could also support frontline clinicians unfamiliar with toxicology by offering immediate insights into likely severity, thereby guiding monitoring intensity and resource allocation. However, successful deployment will require rigorous prospective validation, seamless integration into existing clinical workflows, and user-friendly interfaces that present model explanations clearly enough to foster clinician confidence and accountability.

5 Conclusion

This study shows that machine learning, and particularly gradient-boosted tree models, can provide accurate, interpretable, and clinically relevant predictions of poisoning severity from routinely collected PCC data. By highlighting the most influential variables, our approach not only supports decision-making but also reinforces clinicians' trust in the system. With further validation and expansion, such tools could become a valuable complement to expert judgment, contributing to faster triage, more consistent advice, and ultimately improved patient outcomes in toxicology.

References

[1] F. Mirzakhani, F. Sadoughi, M. Hatami, and A. Amirabadizadeh, "Which model is superior in predicting ICU survival: artificial intelligence versus conventional approaches," *BMC Med. Inf. Decis. Making*, vol. 22, pp. 1–13, Dec. 2022.

[2] H. E. Persson, G. K. Sjöberg, J. A. Haines, and J. P. de Garbino, "Poisoning Severity Score. Grading of Acute Poisoning," *J. Toxicol. Clin. Toxicol.*, Jan. 1998.

[3] K. G. Sam, K. Kondabolu, D. Pati, A. Kamath, G. Pradeep Kumar, and P. G. M. Rao, "Poisoning severity score, APACHE II and GCS: Effective clinical indices for estimating severity and predicting outcome of acute organophosphorus and carbamate poisoning," *J. Forensic Legal Med.*, vol. 16, pp. 239–247, July 2009.

[4] A. F. Sharif, Z. A. Kasemy, R. A. Alshabibi, S. J. Almufleh, F. W. Abousamak, A. A. Alfrayan, M. Alshehri, R. A. Alemies, A. S. Almuhsen, S. N. AlNasser, and K. A. Al-Mulhim, "Prognostic factors in acute poisoning with central nervous system xenobiotics: development of a nomogram predicting risk of intensive care unit admission," *Toxicol. Res.*, vol. 12, pp. 62–75, Feb. 2023.

[5] F. E. Gharbi, N. E. Bèze, K. Jaffal, L. Sutterlin, P. Mora, I. Malissin, N. Deye, S. Voicu, and B. Mégarbane, "Does the ICU Requirement Score allow the poisoned patient to be safely managed without admission to the intensive care unit? – a validation cohort study," *Clin. Toxicol.*, Mar. 2022.

[6] G. N. El-Sarnagawy, A. A. Abdelnoor, A. A. Abuelfadl, and I. H. El-Mehallawi, "Comparison between various scoring systems in predicting the need for intensive care unit admission of acute pesticide-poisoned patients," *Environ. Sci. Pollut. Res.*, vol. 29, pp. 33999–34009, May 2022.

[7] E. S. Schwarz, K. T. Kopec, T. J. Wiegand, P. M. Wax, and J. Brent, "Should We Be Using the Poisoning Severity Score?," *J. Med. Toxicol.*, vol. 13, pp. 135–145, June 2017.

[8] S. Schmoll, E.-C. Heier, R. Böll, T. Zellner, K. Romanek, F. Eyer, C. Rabe, and S. Geith, "Independent validation of the Tanta University Risk Model for intensive care requirement in acutely poisoned adults," *Clin. Toxicol.*, Apr. 2023.

[9] M. Flack, F. Koop, T. Zellner, E.-C. Heier, S. Geith, F. Eyer, C. Rabe, and S. Schmoll, "The Tanta University risk model could help identify patients with acute poisoning who would require intensive care unit level of care," *Clin. Toxicol.*, June 2024.

[10] V. Yousefinejad, B. Moradi, A. M. Baneh, F. Sheikhesmaeili, and A. Babahajian, "Prognostic Factors of Outcome in Methanol Poisoning; an 8-year Retrospective Cross-sectional Study," *Archives of Academic Emergency Medicine*, vol. 8, p. e69, Sept. 2020.

[11] R. Brandenburg, S. Brinkman, N. F. de Keizer, J. Kesecioglu, J. Meulenbelt, and D. W. de Lange, "The need for ICU admission in intoxicated patients: a prediction model," *Clin. Toxicol.*, Jan. 2017.

[12] R. Böll, K. Romanek, S. Schmoll, R. Stich, A. Ott, J. Stenzel, S. Geith, F. Eyer, and C. Rabe, "Independent validation of the ICU requirement score in a cohort of acutely poisoned adults," *Clin. Toxicol.*, July 2018.

[13] F. Martín-Rodríguez, R. López-Izquierdo, M. A. Castro-Villamor, J. L. Martín-Conty, R. M. Herrero-Antón, C. del Pozo-Vegas, D. Guillén-Gil, and A. Dueñas-Laita, "A predictive model for serious adverse events in adults with acute poisoning in prehospital and hospital care," *Aust. Crit. Care*, vol. 34, pp. 209–216, May 2021.

[14] D. M. El-Gharbawy, I. A. Kabbash, and M. M. Ghonem, "A nomogram proposal for early prediction of intensive care unit admission in patients with acute antipsychotic poisoning," *Toxicol. Res.*, vol. 12, no. 5, p. 873, 2023.

[15] S. A. El-Nagdy, A. Y. Alalmie, A. Assiri, A. A. J. Zarbah, A. A. Sobeih, and A. A. Elsheikh, "Proposal of a New Predictive Score for the Requirement of Intensive Care Unit Admission Among Ethanol-Intoxicated Patients," *Alcoholism Treatment Quarterly*, July 2024.

[16] S. Slade, "Case-Based Reasoning: A Research Paradigm," *AIMag.*, vol. 12, p. 42, Mar. 1991.

[17] B. N. Barigou, F. Barigou, C. Benchehida, B. Atmani, and G. Belalem, "The design of a cloud-based clinical decision support system prototype: Management of drugs intoxications in childhood," *International Journal of Healthcare Information Systems and Informatics*, vol. 13, no. 4, pp. 28–48, 2018. Publisher: IGI Global.

[18] B. Mégarbane, "Toxidrome-based Approach to Common Poisonings," *Asia Pac. J. Med. Toxicol.*, vol. 3, pp. 2–12, Mar. 2014.

[19] R. Schmidt, O. Vorobieva, and L. Gierl, "Adaptation problems in therapeutic case-based reasoning systems," in *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 992–999, Springer, 2003. ISSN: 1611-3349 Journal Abbreviation: SpringerLink.

[20] J. M. Alonso, C. Castiello, M. Lucarelli, and C. Mencar, "Modeling interpretable fuzzy rule-based classifiers for medical decision support," in *Data Mining: Concepts, Methodologies, Tools, and Applications*, pp. 1064–1081, IGI Global Scientific Publishing, 2013.

[21] O. Haas, A. Maier, and E. Rothgang, "Rule-Based Models for Risk Estimation and Analysis of In-hospital Mortality in Emergency and Critical Care," *Front. Med.*, vol. 8, p. 785711, Nov. 2021.

[22] J. A. Sanz, M. Galar, A. Jurio, A. Brugos, M. Pagola, and H. Bustince, "Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system," *Appl. Soft Comput.*, vol. 20, pp. 103–111, July 2014.

[23] M. Chary, E. W. Boyer, and M. M. Burns, "Diagnosis of Acute Poisoning using explainable artificial intelligence," *Comput. Biol. Med.*, vol. 134, p. 104469, July 2021.

[24] P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine Learning Techniques for Multimedia*, pp. 21–49, Springer, 2008. Journal Abbreviation: SpringerLink.

[25] M. Rahimi, S. M. Hosseini, S. A. Mohtarami, B. Mostafazadeh, P. E. T. Evini, M. Fathy, A. Kazemi, S. Khani, S. M. Mortazavi, A. Soheili, S. M. Vahabi, and S. Shadnia, "Prediction of acute methanol poisoning prognosis using machine learning techniques," *Toxicology*, vol. 504, p. 153770, May 2024.

[26] S. M. Hosseini, M. Rahimi, M. R. Afrash, P. Ziaeefar, P. Yousefzadeh, S. Pashapour, P. E. T. Evini, B. Mostafazadeh, and S. Shadnia, "Prediction of acute organophosphate poisoning severity using machine learning techniques," *Toxicology*, vol. 486, p. 153431, Mar. 2023.

[27] O. Mehrpour, F. Saeedi, J. Abdollahi, A. Amirabadizadeh, and F. Goss, "The value of machine learning for prognosis prediction of diphenhydramine exposure: National analysis of 50,000 patients in the United States," *Journal of Research in Medical Sciences*, vol. 28, p. 49, June 2023.

[28] H. Li, Z. Liu, W. Sun, T. Li, and X. Dong, "Interpretable machine learning for the prediction of death risk in patients with acute diquat poisoning," *Sci. Rep.*, vol. 14, pp. 1–8, July 2024.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[30] Y.-c. Wu and J.-w. Feng, "Development and Application of Artificial Neural Network," *Wireless Pers. Commun.*, vol. 102, pp. 1645–1656, Sept. 2018.

[31] H.-J. Ye, S.-Y. Liu, H.-R. Cai, Q.-L. Zhou, and D.-C. Zhan, "A Closer Look at Deep Learning Methods on Tabular Datasets," *arXiv*, July 2024.

[32] K. Moulaei, M. R. Afrash, M. Parvin, S. Shadnia, M. Rahimi, B. Mostafazadeh, P. E. T. Evini, B. Sabet, S. M. Vahabi, A. Soheili, M. Fathy, A. Kazemi, S. Khani, S. M. Mortazavi, and S. M. Hosseini, "Explainable artificial intelligence (XAI) for predicting the need for intubation in methanol-poisoned patients: a study comparing deep and machine learning models," *Sci. Rep.*, vol. 14, pp. 1–15, July 2024.

[33] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artif. Intell. Med.*, vol. 133, p. 102423, Nov. 2022.

[34] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *2015 International Conference on Healthcare Informatics*, pp. 21–23, IEEE, 2015. Journal Abbreviation: Published in: 2015 International Conference on Healthcare Informatics.

[35] "Overview of SNOMED CT," Oct. 2016. [Online; accessed 8. Sep. 2025].

[36] O. Mehrpour, F. Saeedi, S. Nakhaee, F. Tavakkoli Khomeini, A. Hadianfar, A. Amirabadizadeh, and C. Hoyte, "Comparison of decision tree with common machine learning models for prediction of biguanide and sulfonylurea poisoning in the United States: an analysis of the National Poison Data System," *BMC Med. Inf. Decis. Making*, vol. 23, pp. 1–11, Dec. 2023.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] P. K. Plunkett, D. G. Byrne, T. Breslin, K. Bennett, and B. Silke, "Increasing wait times predict increasing mortality for emergency medical admissions," *Eur. J. Emerg. Med.*, vol. 18, pp. 192–196, Aug. 2011.

[39] P. Thölke, Y.-J. Mantilla-Ramos, H. Abdelhedi, C. Maschke, A. Dehgan, Y. Harel, A. Kemtur, L. Mekki Berrada, M. Sahraoui, T. Young, A. Bellemare Pépin, C. El Khantour, M. Landry, A. Pascarella, V. Hadid, E. Combrisson, J. O'Byrne, and K. Jerbi, "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data," *Neuroimage*, vol. 277, p. 120253, Aug. 2023.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, June 2002.

[41] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 01–08, IEEE, 2008.

[42] M. M. Patel, C. D. Travers, J. A. Stockwell, E. A. Numur, R. J. Geller, P. P. Kamat, and J. R. Grunwell, "Reducing Childhood Admissions to the PICU for Poisoning (ReCAP2) by Predicting Unnecessary PICU Admissions After Acute Intoxication," *Pediatric Critical Care Medicine*, vol. 19, p. e120, Feb. 2018.

[43] J. J. Rose, M. S. Zhang, J. Pan, M. C. Gauthier, A. F. Pizon, M. I. Saul, and S. M. Nouraie, "Heart-Brain 346-7 Score: the development and validation of a simple mortality prediction score for carbon monoxide poisoning utilizing deep learning," *Clinical toxicology (Philadelphia, Pa.)*, vol. 61, p. 492, July 2023.

[44] O. Mehrpour, F. Saeedi, and C. Hoyte, "Decision tree outcome prediction of acute acetaminophen exposure in the United States: A study of 30,000 cases from the National Poison Data System," *Basic Clin. Pharmacol. Toxicol.*, vol. 130, pp. 191–199, Jan. 2022.

6 Supplementary Material

Variable	Test	Statistic	p value	Significant
Sex	chi2	56.3092	0.0000	True
Circumstance of exposure	chi2	3964.1673	0.0000	True
Place of exposure	chi2	3148.8527	0.0000	True
Type of exposure	chi2	77.7254	0.0000	True
Age	ks	0.0773	0.0000	True
Weight (kg)	ks	0.0329	0.0000	True
Height (cm)	ks	0.1201	0.0256	False
HasSymptoms	chi2	3154.3680	0.0000	True
SNOMED1	ks	0.1494	0.0000	True
SNOMED2	ks	0.1544	0.0000	True
SNOMED3	ks	0.1492	0.0000	True
SNOMED4	ks	0.1193	0.0000	True
SNOMED5	ks	0.0838	0.0000	True
SNOMED6	ks	0.1030	0.0000	True
SNOMED7	ks	0.0874	0.0000	True
SNOMED8	ks	0.0873	0.0000	True
SNOMED9	ks	0.0852	0.0000	True
SNOMED10	ks	0.0895	0.0000	True
SNOMED11	ks	0.0798	0.0000	True
SNOMED12	ks	0.0838	0.0000	True
SNOMED13	ks	0.0871	0.0000	True
SNOMED14	ks	0.0873	0.0000	True
SNOMED15	ks	0.0895	0.0000	True
SNOMED16	ks	0.0822	0.0000	True
SNOMED17	ks	0.0809	0.0000	True
SNOMED18	ks	0.0885	0.0000	True
SNOMED19	ks	0.0854	0.0000	True
SNOMED20	ks	0.0824	0.0000	True
SNOMED21	ks	0.0914	0.0000	True
SNOMED22	ks	0.1030	0.0000	True
SNOMED23	ks	0.0932	0.0000	True
SNOMED24	ks	0.1069	0.0000	True
SNOMED25	ks	0.0878	0.0000	True
Confidence in amount	chi2	3038.5038	0.0000	True
Route of exposure	chi2	1365.5409	0.0000	True
Estimated amount	chi2	1330.8762	0.0000	True
Duration of exposure	chi2	4365.7451	0.0000	True
Agent	chi2	4502.8911	0.0000	True
Delay from exposure to call	ks	0.3128	0.0000	True

Table 8: Statistical differences between the full dataset and the dataset restricted to errors only, for the binary task, for each variable. ks refers to the Kolmogorov-Smirnov test. The alpha value was adjusted to 0.00125 using Bonferroni's method.