Diagnosis of Poisoning Using Probabilistic Logic Networks

Michael Chary
Boston Children's Hospital
michael.chary@childrens.harvard.edu

Edward W Boyer
Brigham & Women's Hospital
eboyer@bwh.harvard.edu

Michele Burns
Boston Children's Hospital
michele.burns@childrens.harvard.edu

Abstract

Medical toxicology is the clinical specialty that treats the toxic effects of substances, be it an overdose, a medication error, or a scorpion sting. The volume of toxicological knowledge has, as with other medical specialties, outstripped the ability of the individual clinician to master and stay current with it. The application of machine learning techniques to medical toxicology is challenging because initial treatment decisions are often based on a few pieces of textual data and rely heavily on prior knowledge. Moreover, ML techniques often do not represent knowledge in a way that is transparent for the physician, raising barriers to usability. Rule-based systems and decision tree learning are more transparent approaches, but often generalize poorly and require expert curation to implement and maintain. Here, we construct a probabilistic logic network to represent a portion of the knowledge base of a medical toxicologist. Our approach transparently mimics the knowledge representation and clinical decision-making of practicing clinicians. The software, dubbed Tak*, performs comparably to humans on straightforward cases and intermediate difficulty cases, but is outperformed by humans on challenging clinical cases.* Tak *outperforms a decision tree classifier at all levels of difficulty. Probabilistic logic provides one form of explainable artificial intelligence that may be more acceptable for use in healthcare, if it can achieve acceptable levels of performance.*

1. Introduction

The scope of biomedical knowledge is too vast and rate of increase of that knowledge too rapid for an individual physician to bring all relevant knowledge to bear on the diagnosis and treatment of an illness. Machine learning and artificial intelligence (ML/AI) approaches, trained on data sets larger than any physician could encounter in training, can outperform physicians on specific tasks, such as predicting the likelihood of response to a chemotherapeutic treatment [1] or diagnosing pneumonia from a chest X-ray [2]. They perform much worse than physicians in poorly-defined tasks such as constructing the differential diagnosis, a list of diagnoses ranked by the likelihood of explaining a patient's current condition.

A barrier to integrating ML/AI into healthcare is the difference between how ML/AI and physicians arrive at a diagnosis. Many current ML/AI approaches look for quantitative patterns across large data sets, but ignore prior knowledge and cannot explain their reasoning in terms of biomedical knowledge. Pretrained models and transfer learning incorporate statistical relationships from prior data, but do not explicitly represent those relationships in terms that domain experts can readily interpret. Humans frequently use symbolic reasoning to describe complex systems, but ML/AI approaches rarely do. Reliability plots and counterfactual reasoning can explain algorithms that classify pictures of biopsies [3], but it is not clear how these approaches could apply to textual data. Probabilistic logic provides a way to combine statistical learning with symbolic reasoning, to combine machine capacity with human intuition. Our goal was to develop an approach whose ratiocination was transparent to physicians, the intended user.

The organization of this paper is as follows. Section 1 presents a primer of the relevant clinical concepts (Section 1.1), literature review (Section 1.2), a description of probabilistic logic networks (Section 1.3), and alternative approaches that have been used in medical diagnosis (Section 1.4). Section 2 describes the construction of our probabilistic logic network. Section 3 presents the results, and Section 4 the conclusions.

1.1. The Diagnosis and Treatment of a Poisoned Patient

The goal of this section is to briefly describe the clinical reality of treating poisoned patients, provide

URI: https://hdl.handle.net/10125/71050
978-0-9981331-4-0

readers with the minimal necessary clinical context, and compare the diagnosis of a poisoned patient to multinomial classification. We do this to explain why we chose toxidromes as features, the errors that are acceptable to the medical toxicologist, and the relevant benchmarks for performance.

The diagnosis and treatment of a poisoned patient begins with a rapid determination of whether the patient requires immediate intervention to prevent death. This determination is usually made at the patient's bedside by a physical examination and, if the patient is conscious and coherent, a brief discussion with the patient. Consider a patient with progressively slowing breathing. This raises concern for the use of, among other things, a life-threatening opioid overdose. If the patient's breathing continues to slow and no other explanation is more likely, an opioid ingestion is assumed because the drug effect needs to be immediately reversed to prevent death from lack of oxygen. Opioids can slow breathing within minutes of ingestion, but its metabolites are not detected in urine for hours. Measuring blood concentrations of opioids often requires specialized equipment and the results are not available for days. Serum or urine drug concentrations, in general, are not available quickly enough to inform life-saving interventions.

The goal of bedside evaluation of the poisoned patient is to identify reasons to administer time-sensitive medications to prevent death. After treatment, the patient is re-evaluated for a response. The patient's response to the intervention guides future interventions and refinements in the diagnosis, until the patient is stabilized. A unique aspect of toxicology is that some treatments are simultaneously diagnostic and therapeutic. For example, administration of the medication naloxone selectively reverses slowing of breathing due to opioids (also called narcotics). If a patient responds to naloxone, that patient's breathing is restored and posterior probability that an opioid caused their respiratory has increased.

The bedside evaluation a medical toxicologist performs resembles feature extraction and then multinomial classification with a complex loss function. Table 1 displays the 6 canonical toxidromes, the features physicians extracts from bedside evaluation, and the feature values for each toxidrome. The sedative-hypnotic toxidrome is the least lethal toxidrome. Misclassifying a sedative-hypnotic toxidrome as another is less costly than misclassifying another toxidrome as sedative-hypnotic. From this lens, a toxidrome is a set of ranges of values of features that define decision regions for class membership. The word *toxidrome* refers to a decision region. Toxidromes are intended to accurately identify severe poisonings that will respond to treatment, but may misclassify milder poisonings. This misclassification is acceptable clinically because mild poisonings, in general, do not require any specific treatment.

	HR	BP	Pup	Sec	Temp	RR	MS
Anticholinergic	⇑		●	⇓	⇑		D
Cholinergic	⇓		•	⇑		⇓	S
Opioid			•			⇓	S
Sedative-Hypnotic							S
Serotonin Toxicity	⇑	⇑			⇑		A
Sympathomimetic	⇑	⇑	●		⇑	⇑	A

Table 1. **Six canonical toxidromes.** HR, heart rate; BP, blood pressure; Pup, pupil diameter, size of bullet represents increased or decreased pupil diameters; Sec, secretions; Temp, temperature; RR, respiratory rate; MS, mental status; D, delirious, S, sedated; A, agitated Empty cell indicates expectation of no abnormality for that sign.

Some features are costly to observe or are unobservable in some patients. A patient may not make urine owing to kidney failure. Some drug level concentrations are only informative at certain intervals after ingestion, reflecting ongoing partitioning between the bloodstream and other parts of the body. A patient may not display all the signs of a toxidrome. Drug effects vary with age and prior usage. For example, the opioid buprenorphine (trade name Suboxone) is more likely to slow breathing in children than adults. The substance may not be completely absorbed at the time of the first bedside evaluation. A patient's toxidrome may change over time if the patient consumed life-threatening amounts of many drugs with different rates of absorption or metabolism, for example heroin and cocaine. Genetic polymorphisms may accelerate or slow metabolism. A patient may have taken substances that interact with prescribed medication.

The inter-rater reliability among toxicologists in using toxidromes to diagnose poisonings has not been systematically studied. The need for such a study is acknowledged[4]. The data provided here are, to the authors' knowledge, the first attempt to establish such a benchmark.

The names of the toxidromes reflect the biochemical pathways excessively activated or blocked by classes of drugs. The anticholinergic

toxidrome results from blockade of the acetylcholine receptor family; cholinergic toxidrome from activation of the acetylcholine receptor family; opioid toxidrome from activation of the μ opioid receptors; sedative-hypnotic toxidrome from activation of the GABA (γ-amino butyric acid) receptors or blockade of glutamate receptors, and the sympathomimetic toxidrome from activation at adrenaline or noradrenaline receptors [5]. Serotonin toxicity is thought to result from excess activation at the 5-HT_{2A} receptor[6].

1.2. Review of Relevant Studies

Machine learning techniques (*e.g.* Bayesian classifiers, neural networks, decision trees) have been applied to medical diagnosis[7]. Here we review the application of probabilistic logic networks that analyze text to medical diagnosis. We exclude algorithms that use images, despite their success in radiology and pathology. Support vector machines have been productively applied to identify poisonous mushrooms [8] and plants [9, 10] from images. But, image data are often not available to the toxicologist.

To the authors' knowledge there have been no prior publications on the application of probabilistic logic networks to the diagnosis of the poisoned patient. The Chemical Hazards and Emergency Medical Management branch (CHEMM) of the US Department of Health and Human Services developed the CHEMM Intelligent Syndromes Tool. This tool is available via a web-based interface, but no application programming interface or scalable endpoint is provided. There are no publications describing its implementation, but it appears to be based on FALCON[11], a deterministic decision tree to co-ordinate responses against attacks with chemical weapons.

A combination of Bayesian networks and an ontology has been used to diagnose osteoporosis, achieving a 72% accuracy [12]. The diagnosis of osteoporosis is made based on the value of one parameter, a person's average bone density. A person is diagnosed with osteoporosis if his or her average bone density is 2.5σ below the reference distribution for the general population. Poisonings, however, are diagnosed by the presence of overlapping categorical features. A Markov logic network has been implemented for diagnosis of medical conditions from Chinese-language medical records[13], but there was no assessment of its performance. The construction of a fuzzy Bayesian network for medical diagnosis has also been proposed but its performance not assessed [14], as has a case-based learning approach to represent clinical reasoning about breast cancer [15].

1.3. Probabilistic Logic Networks

Probabilistic logic networks (PLNs) aim to represent knowledge about the world and allow inference under uncertainty using a combination of predicate logic, symbolic reasoning, and statistical inference[16]. A PLN consists of a set of pairs of a probability and a logical statement.

$$0.4 \quad \text{somnolent}\,(x) \Rightarrow \text{sedative\_hypnotic}\,(x) \quad (1)$$

Equation (1) represents the concept that in 40% of possible worlds if a patient is somnolent (excessively sleepy and lethargic) then the patient may be poisoned by a medication from the sedative/hypnotic class.

Software that combined rules and probabilities was developed for medicine as early as the 1970's, *e.g.* MYCIN[17]. Neural networks eclipsed rule-based systems because they could operate with inexact matching, performed more accurately, and scaled more easily and rapidly. The adoption of neural networks by physicians is limited, however, owing, in part, clinicians' reluctance to interact with something that "doesn't speak my language"[18].

We use ProbLog, a PLN implementation that treats logical statements as random variables [19]. The fraction associated with each statement represents the fraction of words in which the logical statement is true. The statement is assumed to be false in all other worlds. This is sometimes termed *distributional semantics* [20]. Equation 2 shows how the probability associated with a query Q is related to the logical rules and their associated probabilities. The sum ranges over all worlds in which the supplied facts, F, and stated rules, R, imply that the query Q is true. The products range over all worlds. The first product calculates the joint probability of the supplied facts being true. The second product calculates the joint probability of other facts being false. In our case, R is the set of rules specifying the relationships between clinical findings and toxidromes. The supplied facts, F, correspond to clinical findings. The query Q is for each toxidrome. The directive is to find the query (toxidrome) that maximizes Equation 2.

$$P\,(Q) = \sum_{F \cup R \models Q} \prod_{f \in F} p\,(f) \prod_{f \notin F} 1 - p\,(f) \quad (2)$$

Rule-based systems need experts to create and curate the rules as well as to adapt the rules to include new knowledge or apply the system to unfamiliar types of data. This need for expert curation may limit the speed of development. In the context of developing applications that mimic medical thought, involving domain experts may increase adoption. For a more complete introduction to probabilistic logic we refer the reader to [21] and for ProbLog to [22].

An advantage of rule-based approaches over other machine learning approaches for medicine is that rule-based approaches do not require large training data sets. Many specialties within medicine diagnose and treat rare diseases for which data sets large enough to explore all methods of diagnosis and treatment are unlikely to exist. Rules can be a distillation of the received knowledge of a field, or a combination of this distillation and relationships inferred from large data sets, increasing model transparency to the clinician.

1.4. Alternative Approaches

Decision tree (DT) learning provides a competitive alternative to PLNs. DT classifiers can predict the risk of breast cancer[23], heart disease[24], and diagnose diabetes[25]. DTs are robust against collinearity. This is important in toxicology where poisonings share overlapping features. An elevated heart rate can be seen in 3 toxidromes (anticholinergic, sympathomimetic, and serotonin toxicity). In 2 of those toxidromes elevated blood pressure also occurs.

Similar to PLNs, DTs are straightforward to understand and interpret. It is important that algorithms used in the care of patients be able to explain their information processing in ways that are explainable to and agree with the domain-specific knowledge of physicians. Otherwise, physicians may be less willing to incorporate these algorithms into their medical decision making. DTs and PLNs require less training data than neural networks. This is advantageous for medical applications, where curated data is often tiny. All studies discussed in Section 1.2 were developed on 150 or fewer patient presentations.

A limitation of DTs is their tendency to overfit. In medicine the goal of accurately diagnosing (classifying) is often balanced against the heuristic of recognizing a misdiagnosis (misclassification). PLNs avoid this overfitting because they are not trying to minimize the heterogeneity of classes.

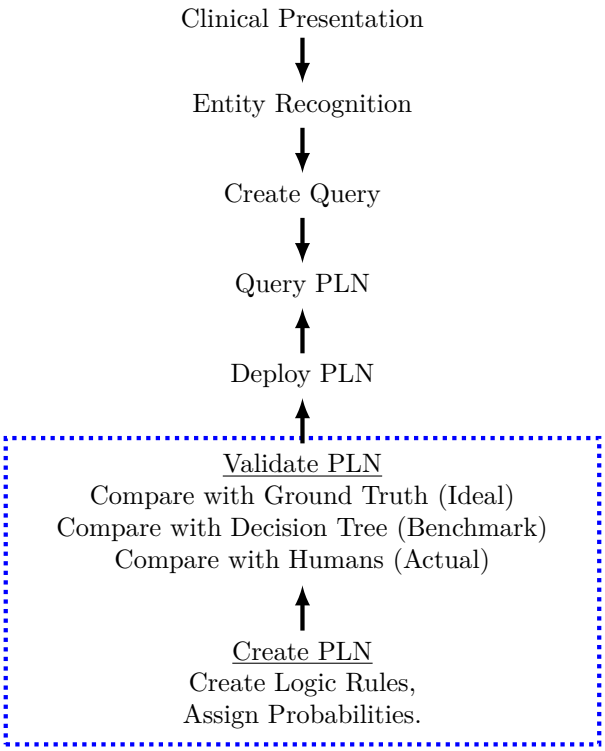

Figure 1. **Flowchart of project and future directions.** Dotted rectangle indicates scope of current paper. PLN, probabilistic logic network. Ideal (performance); Actual (performance).

2. Methods

Figure 1 shows our general approach. The notation is described in more detail below.

2.1. Knowledge Representation

We created 34 probabilistic logic rules based on the consensus of three medical toxicologists. We named these rules and the underlying implementation in ProbLog, *Tak*. These rules described the medical knowledge of the features used to diagnose each toxidrome. We restricted ourselves to developing rules that described features that could be observed during one evaluation at a patient's bedside without laboratory testing. We did this to assess our approach in the most time-sensitive aspect of medical toxicology. The rules were constructed as sets of predicates as follows. We treated each finding elicited by the toxicologist as a predicate. For example, the predicate `salivation(X, increased)` is true if patient

```
0.10::salivation(X,decreased);
0.10::salivation(X,increased);
0.80::salivation(X,usual).
```

Listing 1. **Example delineation of prior probabilities across cardinal sign with annotated disjunction.** Number preceding each goal denotes number of worlds in which that goal is satisfied. Numbers sum to one across values.

demonstrates increased salivation. We constructed all predicates to have two slots, the patient and the value of the feature, usually {present|absent} or {increased|normal|decreased}. These values reflect a discretization of underlying continuous variables, a common pattern of communication with information compression between physicians. For example, the respiratory rate is quantified as the number of breaths a patient takes each minute. Physicians, more commonly, describe a patient as having an increased or decreased respiratory rate, implicitly referring to an expected normal value, rather than stating the absolute number. We modeled medical decision making at the level of categorical variable to reflect the nature of communication among physicians. Anecdotally, physicians prefer to communicate in a mix of categorical and continuous variables, favoring categorical variables as complexity increases.

This approach created 17 rules that describe the prior distribution of feature values (example in Listing 1), 10 rules that describe the posterior probability of a toxidrome given clinical findings (example in Listing 2), and 7 that describe the sufficient conditions for each toxidrome (example in Listing 3). These rules were written in ProbLog, a Prolog extension for probabilistic logic[22].

In Listing 1, the number before the colon represents the fraction of worlds in which the following predicate is true. Listing 2 demonstrates assigning the relative chance of one toxidrome over another as a consequence of an entity manifesting a symptom. The function `mentalStatus(X, agitated)` is true if patient X is agitated. The function `hasToxidrome(X,Y)` is true if patient X manifests toxidrome Y.

The relative probabilities across rules were chosen to reflect the perceived relative prevalence of each value. We used the most recent annual report from American Association of Poison Control Centers on the relative prevalence of each poisoning in the US to estimate the prior probability of each toxidrome. Probability distributions do

```
4*P::hasToxidrome(X, sympathomimetic);
P::hasToxidrome(X, serotonergic) :-
mentalStatus(X, agitated), P is 0.2.
```

Listing 2. **Example calculation of posterior probability of toxidromes given value of cardinal sign.** Expression preceding each goal in disjunction (sequence of statements separated by semicolons) is evaluated when conditions of goal are satisfied.

```
hasToxidrome(X, cholinergic) :-
salivation(X, increased),
urination(X, increased),
pupilDiameter(X, small).
```

Listing 3. **Example Expression of Diagnosis of Toxidrome as Prolog Goal.**

not exist for most features. The prevalence of, for example, hypersalivation in the general population is not known. Nor is it known that a patient is exactly four times more likely to suffer from a sympathomimetic toxidrome as opposed to serotonin toxicity if the patient becomes agitated after an unknown ingestion. The clinical findings from excess serotonin activation are classically called serotonin syndrome or serotonin toxicity, but the term is used equivalently to a toxidrome.

2.2. Data Set

We generated 300 simulated toxidrome presentations following Algorithm 1 from the list of 6 toxidromes by randomly sampling from a uniform distribution with replacement. For each toxidrome we created a presentation of 5 signs, $5-k$ related to the toxidrome and k unrelated to the toxidrome. The parameter k, which we term *difficulty*, allows one to simulate the variability of clinical presentation. A difficulty of 0 simulates an unequivocal presentation. A difficulty of 2 simulates a mixed picture, as might result from the ingestion of many substances with conflicting effects. Table 2 shows the distribution of simulated presentations across intended toxidromes and difficulty. This approach mimics the early stages of the training of a medical toxicologist, wherein they are exposed to simple stylized cases. To the author's knowledge, this is the first data set created to evaluate the inter-rater reliability of diagnosis in medical toxicology.

	Difficulty		
	0	1	2
Anticholinergic	14	14	14
Cholinergic	19	26	16
Opioid	25	11	13
Sedative-Hypnotic	17	13	24
Serotonin Toxicity	10	21	20
Sympathomimetic	15	15	15

Table 2. **Distribution of Simulated Presentations.**

Algorithm 1 Generation of simulated toxidrome

Precondition: $n \leftarrow 5$
▷ Maximum number of signs per presentation
Precondition: $0 \leq k \leq n$ ▷ difficulty
Precondition: $\{t\} \Leftarrow$ {sympathomimetic, anticholinergic, cholinergic, sedative_hypnotic, opioid, serotonin_toxicity} ▷ toxidromes
Precondition: $\{t, s, v\} \Leftarrow \{(t_i, s_{ij}, v_{ijk})\}$
▷ classic values for each sign in a toxidrome

Postcondition: $\{p\} \Leftarrow \{s_k, v_k\}$
▷ set of sign, value pairs

$t_i \Leftarrow$ random.choice $(\{t\})$ ▷ intended toxidrome
$\hat{t}_i \Leftarrow$ random.choice $(\{t\} - t_i)$ $i \neq j$
▷ distractor toxidrome
presentation $\leftarrow \{\}$
while generated = false **do**
chose $(5-k)$ $\{s, v\}$ pairs from $\{(t = t_i, s_{ij}, v_{ijk})\}$
presentation $\leftarrow$ pairs
chose k $\{s, v\}$ pairs from $\{\left(t = \hat{t}_i, s_{ij}, v_{ijk}\right)\}$
presentation $\leftarrow$ pairs
end while

2.3. Evaluation

We compared the performance of *Tak*, the probabilistic logic network, with expert consensus, a decision tree, and ideal performance (recovery of the true labels generated by Algorithm 1). We assessed actual and ideal performance by comparing *Tak*'s outputs with a data set labeled by manual curation. We presented the same cases to *Tak* and human raters TC (Takuyo Chiba, M.D.) and AB (Alexander Barbuto, M.D.), two medical toxicologists. *Tak* assigned the most likely rating to each toxidrome. TC and AB labeled each presentation with the toxidrome they felt most accurately described the presentation. We omitted a presentation if either rater could not assign a toxidrome ($n = 18$), decreasing the number of cases from 300 to 282. Our evaluation of actual performance is limited by omitting these cases.

We use the term *inferred toxidrome* to denote the toxidrome that *Tak*, the decision tree, or the human raters inferred from the case presentation. We use the term *intended toxidrome* to denote the actual toxidrome associated with each case presentation, *i.e.* the true label. The term *human raters* refers to two medical toxicologists, AB and TC, who independently reviewed each case presentation and inferred the toxidrome. We quantified inter-rater reliability using a multinomial extension of Cohen's κ. Cohen's κ ranges between 0 and 1 where 0 indicates the level of agreement expected by chance and 1 indicates perfect agreement.

We used the inter-rater reliability between *Tak*'s inferred toxidrome and the intended toxidrome as a measure of best performance and between *Tak*'s inferred toxidrome and the consensus of the human raters as a measure of actual performance. We took the inter-rater reliability between the consensus of the human raters and the generated labels as a benchmark for actual performance.

To provide a machine learning benchmark we trained a decision tree classifier[26] on the same cases. We trained one decision tree classifier for each level of difficulty. This approach is likely to lead to overfitting, but also will overestimate the decision tree's performance, providing a more stringent benchmark against which to evaluate *Tak*. We used the *sklearn* implementation, DecisionTreeClassifier, with the maximum depth set to 3.

3. Results

3.1. Ideal Performance (*Tak* vs. Intended Toxidrome)

Tables 3 - 5 provide the confusion matrices between *Tak*'s inferences and the intended toxidromes. The inter-rater reliabilities were $\kappa = 0.8554$, $\kappa = 0.5614$, and $\kappa = 0.2904$ for difficulty levels 0,1, and 2, respectively. For comparison, the inter-rater reliabilities between the consensus of human raters and the intended toxidromes were $\kappa = 0.9878$, $\kappa = 0.7818$, and $\kappa = 0.2718$ for difficulty 0,1, and 2, respectively.

As the difficulty increased, the difference between *Tak*'s accuracy and the human consensus accuracy decreased. In simpler presentations, Tak confused the sedative-hypnotic toxidrome with the opioid toxidrome in $1/17 = 5.8\%$ of presentations and with the anticholinergic toxidrome in $1/17 =$

$\kappa = 0.8554$

Intended Toxidrome \ Intended	Anticholinergic	Cholinergic	Opioid	Sedative Hypnotic	Serotonin Toxicity	Sympathomimetic
Anticholinergic	14	0	0	0	1	0
Cholinergic	0	19	0	0	0	0
Opioid	0	0	25	0	0	0
Sedative Hypnotic	1	0	1	15	0	0
Serotonin Toxicity	0	0	0	0	10	0
Sympathomimetic	0	0	0	0		15

Table 3. **Confusion matrix between intended toxidromes and *Tak* for difficulty** 0**.** Axis shows rater. Axis label, mostly likely toxidrome; Number, color in each grid, the number of diagnoses; κ denotes Cohen's κ.

$\kappa = 0.5452$

Intended Toxidrome \ Tak	Anticholinergic	Cholinergic	Opioid	Sedative Hypnotic	Serotonin Toxicity	Sympathomimetic
Anticholinergic	7	2	0	4	1	0
Cholinergic	0	26	0	0	0	0
Opioid	0	3	8	0	0	0
Sedative Hypnotic	1	0	1	11	0	0
Serotonin Toxicity	0	0	0	0	21	0
Sympathomimetic	0	1	0	4	1	9

Table 4. **Confusion matrix between intended toxidromes and *Tak*'s predicted toxidromes for difficulty** 1**.** Axis shows rater. Axis label, mostly likely toxidrome; Number, color in each grid, the number of diagnoses; κ denotes Cohen's κ.

5.8% of presentations. As difficulty increased *Tak* developed difficulty distinguishing among the anticholinergic, cholinergic, and sedative-hypnotic toxidromes as well as between the opioid and cholinerigc toxidromes.

These misclassification errors arise because the features used to describe toxidromes are collinear and features that are pairwise independent do not have values for all toxidromes. For example, in a poisoned patient the heart rate and blood pressure usually move in tandem, both rising or both falling. This collinearity is such a hallmark of poisoned patients that its absence can prompt medical toxicologist to consider nontoxicological causes of the patient's illness. The ranges of values a feature takes on may not allow discrimination between toxidromes. The mental status of someone with the sympathomimetic toxidrome or serotonin toxicity can be classified as agitated. Section 3.4 discusses this in more detail.

3.2. Usual Performance (*Tak* vs. Human Consensus)

Our benchmark for usual performance was the inter-rater reliability between the consensus of human raters and the intended toxidromes. The inter-rater reliability between the consensus of the human raters and the labels predicted by *Tak* was $\kappa = 0.8432$, $\kappa = 0.4396$, and $\kappa = 0.3331$, for difficulties 0,1, and 2, respectively. Table 6 shows the confusion matrix for difficulty=0 presentations.

The decrease in inter-rater reliability for more difficult cases arose, in part, from the introduction of feature values that were equally predictive of more than one toxidrome. For example, an elevated heart rate can be a sign of the anticholinergic or sympathomimetic toxidrome and, indeed, it can be difficult for clinicians to distinguish these two processes without further information. Similarly, small pupils can be a sign of the cholinergic or opioid toxidrome. Of the 11 cases of opioid toxidromes, AB classified 2 as cholinergic that TC classified as opioid. TC classified 1 as cholinergic that AB classified as opioid. The opioid and cholinergic toxidromes overlap clinically. Both manifest as with constricted pupils, somnolence, and slowed respiratory rate (confusion matrix not shown). To the author's knowledge this is the first published attempt to quantify the inter-rater reliability of medical toxicologists on any data set.

3.3. Benchmark (Decision Tree vs. Intended Toxidromes)

To compare *Tak*'s performance against other machine learning approaches, we calculated the inter-rater reliability between the ground truth labels and a decision tree classifier. In straightforward presentations (difficulty, 0) *Tak* outperformed the decision tree ($\kappa_{\mathrm{DT}} = 0.6144$ vs $\kappa_{Tak} = .8554$). In intermediate difficulty

Intended Toxidrome \ Tak	Anticholinergic	Cholinergic	Opioid	Sedative Hypnotic	Serotonin Toxicity	Sympathomimetic
Anticholinergic	7	2	0	3	0	2
Cholinergic	0	13	2	0	1	0
Opioid	0	3	4	3	1	2
Sedative Hypnotic	2	5	4	12	0	1
Serotonin Toxicity	1	3	2	4	8	2
Sympathomimetic	5	0	2	0	1	5

$\kappa = 0.3065$

Table 5. **Confusion matrix between intended toxidromes and *Tak*'s predicted toxidromes for difficulty 2.** Axis shows rater. Axis label, mostly likely toxidrome; Number, color in each grid, the number of diagnoses; κ denotes Cohen's κ.

presentations *Tak* outperformed the decision tree ($\kappa_{\text{DT}} = 0.3527$ vs $\kappa_{Tak} = .5614$). In complex presentations (difficulty=2) *Tak* performed comparably to the decision tree ($\kappa_{\text{DT}} = 0.2622$ vs $\kappa_{Tak} = .2904$). This benchmark was designed to favor the decision tree. A separate decision tree was fitted for each class of difficulty. The trees were not averaged nor was any random forest method used.

Intended Toxidrome \ Human Raters	Anticholinergic	Cholinergic	Opioid	Sedative Hypnotic	Serotonin Toxicity	Sympathomimetic
Anticholinergic	14	0	0	0	0	0
Cholinergic	0	18	0	0	0	0
Opioid	0	2	25	3	0	0
Sedative Hypnotic	0	0	0	4	0	2
Serotonin Toxicity	0	0	0	0	12	0
Sympathomimetic	0	0	0	0	0	15

$\kappa = 0.8432$

Table 6. **Confusion matrix between consensus of human raters and predicted diagnoses for difficulty 0.** Axis shows rater. Axis label, mostly likely toxidrome; Number, color in each grid, the number of diagnoses; κ denotes Cohen's κ.

3.4. Evaluation of Errors

As presentation difficulty increased *Tak* and the human raters both decreased in accuracy. This decrease in accuracy reflects the construction of the synthetic data set and the limits of semantic resolution of toxidromes. The synthetic data were constructed to have three levels of difficulty, corresponding to clinical reality. Some patients may ingest or be exposed to a large amount of one substance leading to an unequivocal presentations. Others may ingest or be exposed to many substances with a variety of stimulating and sedating effects, the balance of which shifts over time as the chemicals are distributed throughout the body and metabolized at different rates.

Section 3.1 described how the features used to discriminate toxidromes create overlapping decision regions. The authors could find no published analysis of the discriminative limits of toxidromes, but it stands to reason that 6 features cannot accurately classify into 6 categories if they are collinear and do not have values for all categories.

Tak confused the anticholinergic and sympathomimetic toxidromes and cholinergic, opioid, and sedative hypnotic toxidromes. This mimics error patterns of medical toxicologists. The anticholinergic and sympathomimetic toxidromes share features (increased heart rate, increased blood pressure, and agitation). The cholinergic, opioid, and sedative-hypnotic toxidromes share features (sedation, and the cholinergic and opioid toxidromes share slowed breathing and small pupils). *Tak* could distinguish serotonergic toxicity from the anticholinergic and cholinergic toxidromes because serotonergic toxicity has unique features. Future work can explore the sensitivity of classification to each rule.

4. Conclusions

The goal of this study was to present a novel application of using probabilistic logic to model medical knowledge. We derived probabilistic logic rules from expert consensus, constructing a probabilistic logic network, *Tak*. We evaluated *Tak*'s performance on a synthetic data set and compared its performance against the consensus of domain experts and a decision tree classifier.

Figure 2 summarizes the peak, actual, and benchmark performances. *Tak*'s peak and actual performance were comparable (Ground Truth vs *Tak* and (Human vs *Tak*). As case complexity

		$\kappa = 0.3331$					
		Inferred Toxidrome					
		Anticholinergic	Cholinergic	Opioid	Sedative Hypnotic	Serotonin Toxicity	Sympathomimetic
Intended Toxidrome	Anticholinergic	5	1	3	3	0	0
	Cholinergic	0	12	3	2	0	0
	Opioid	0	1	7	3	0	0
	Sedative Hypnotic	0	0	0	11	0	0
	Serotonin Toxicity	0	0	0	2	11	0
	Sympathomimetic	1	0	0	0	2	9

Table 7. **Confusion matrix between consensus of human raters and *Tak* diagnoses for difficulty 2.** Axis shows rater. Axis label, mostly likely toxidrome; Number, color in each grid, the number of diagnoses; κ denotes Cohen's κ.

increased *Tak*'s accuracy decreased, as did the accuracy of human raters. In the most difficult cases, *Tak*'s accuracy approached that of the decision tree classifier, approximately half of human performance. In all cases, *Tak's* performance was better than chance ($\kappa = 0$) and the decision tree, our benchmark for current approaches.

The misclassification errors made by *Tak* resemble the misclassification errors by humans, for example confusion between the opioid and cholinergic toxidromes and between the sympathomimetic toxidrome and serotonin toxicity. *Tak*'s performance is As it currently stands *Tak*'s error rate needs improvement before clinical use. Even if *Tak* never reaches accuracy comparable to human physicians across all levels of difficulty, it may still have a role in reducing physician error by providing a near real-time "second opinion" and generating a signal when *Tak*'s diagnosis and the physician's diagnosis differ. This could help mitigate physician error due to fatigue. In addition, *Tak* could be used to automate the processing of more routine cases, freeing up physician time to deal with more complex cases.

The most significant limitation of this paper is that we evaluated our approach on synthetic rather than actual clinical data. We used synthetic data because no such clinical data set currently exists. We took steps to generate realistic data, creating cases at three levels of complexity to capture some of the heterogeneity of clinical data. We reviewed all cases with 2 medical toxicologists and excluded 18 cases they felt were not clinically plausible. Our data set of 300 cases is not large enough to fully explore the feature space, which would require $\left(3^5\right) \cdot 4^5$ unique cases. Many regions of the full feature space are not medically plausible, but their evaluation could provide further insight into *Tak*. The full value of our approach in helping physicians treat poisoned patients will not be known until our approach can be evaluated on actual clinical data.

For most variables empiric probability distributions were not available. This renders the absolute values of the calculated posterior probabilities uninformative even if the relative magnitude is still informative. A limitation of using categorical variables is that another processing layer will be needed between entity recognition and querying *Tak* to convert continuous variables into categorical variables.

Physicians must trust an AI-based system to include it in their evaluation and treatment of patients. An algorithm can earn that trust through proficiency on complex cases and transparency. *Tak* demonstrates transparent clinical reasoning. This transparency, if preserved in more accurate models, may remove barriers to the use of AI approaches in clinical decision making. Even if a more detailed analysis of the limits of PLNs suggests a unimprovably poor performance on complex cases, a transparent AI-system may be useful by automating aspects routine cases and in doing so freeing up expert time for more complicate cases. This is similar to the use of nurse practitioners and physician assistants in some specialties.

The main contribution of this paper is the demonstration that probabilistic logic networks can model toxicologic knowledge in a way that transparently mimics physician thought. This paper is also the first, to the authors' knowledge, to quantify the inter-rater reliability of physicians in diagnosing poisoning. Yet another contribution is the development of a data set that unsupervised or weakly supervised techniques could use to explore other knowledge representations in this domain.

References

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[2] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

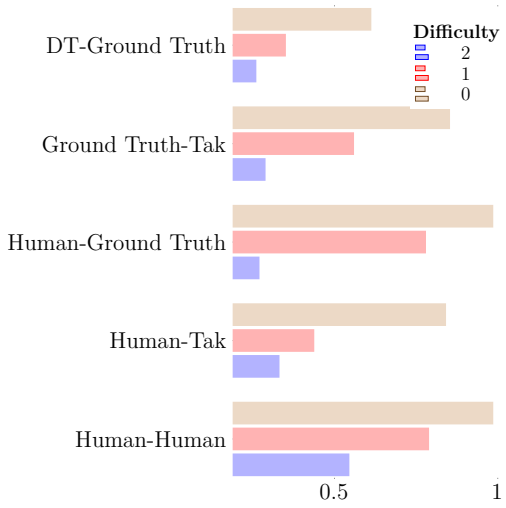

Figure 2. **Human and Computer Agreement** Y-axis denotes Cohen's κ; X-axis, raters; Hue, difficulty of presentation, *i.e.* k

[3] J. J. Thiagarajan, P. Sattigeri, D. Rajan, and B. Venkatesh, "Calibrating healthcare ai: Towards reliable and interpretable deep predictive models," *arXiv preprint arXiv:2004.14480*, 2020.

[4] J. Larsen, M. B. Mycyk, and T. M. Thompson, "Reviewing the record: Medical record reviews for medical toxicology research," 2018.

[5] C. P. Holstege and H. A. Borek, "Toxidromes," *Critical care clinics*, vol. 28, no. 4, pp. 479–498, 2012.

[6] E. W. Boyer and M. Shannon, "The serotonin syndrome," *New England Journal of Medicine*, vol. 352, no. 11, pp. 1112–1120, 2005.

[7] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[8] A. Wibowo, Y. Rahayu, A. Riyanto, and T. Hidayatulloh, "Classification algorithm for edible mushroom identification," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, pp. 250–253, IEEE, 2018.

[9] D. S. Prasvita and Y. Herdiyeni, "Medleaf: mobile application for medicinal plant identification based on leaf image," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 3, no. 2, pp. 5–8, 2013.

[10] S. Sharma and C. Gupta, "A review of plant recognition methods and algorithms," *International Journal of Innovative Research in Advanced Engineering*, vol. 2, no. 6, pp. 111–116, 2015.

[11] S. P. Frysinger, M. L. Deaton, A. G. Gonzalo, A. M. VanHorn, and M. A. Kirk, "The falcon decision support system: Preparing communities for weapons of opportunity," *Environmental Modelling & Software*, vol. 22, no. 4, pp. 431–435, 2007.

[12] P. Agarwal, R. Verma, and A. Mallik, "Ontology based disease diagnosis system with probabilistic inference," in *2016 1st India International Conference on Information Processing (IICIP)*, pp. 1–5, IEEE, 2016.

[13] J. Jiang, X. Li, C. Zhao, Y. Guan, and Q. Yu, "Learning and inference in knowledge-based probabilistic model for medical diagnosis," *Knowledge-Based Systems*, vol. 138, pp. 58–68, 2017.

[14] V. Zarikas, E. Papageorgiou, D. Pernebayeva, and N. Tursynbek, "Medical decision support tool from a fuzzy-rules driven bayesian network.," in *ICAART (2)*, pp. 539–549, 2018.

[15] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial intelligence in medicine*, vol. 94, pp. 42–53, 2019.

[16] B. Goertzel, M. Iklé, I. F. Goertzel, and A. Heljakka, *Probabilistic logic networks: A comprehensive framework for uncertain inference*. Springer Science & Business Media, 2008.

[17] E. H. Shortliffe, "Mycin: a rule-based computer program for advising physicians regarding antimicrobial therapy selection.," tech. rep., STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE, 1974.

[18] C. D. Naylor, "On the prospects for a (deep) learning health care system," *Jama*, vol. 320, no. 11, pp. 1099–1100, 2018.

[19] L. De Raedt, A. Kimmig, and H. Toivonen, "Problog: A probabilistic prolog and its application in link discovery.," in *IJCAI*, vol. 7, pp. 2462–2467, Hyderabad, 2007.

[20] T. Sato, "A statistical learning method for logic programs with distribution semantics," in *IN PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON LOGIC PROGRAMMING (ICLP'95*, Citeseer, 1995.

[21] A. Kimmig, B. Demoen, L. De Raedt, V. S. Costa, and R. Rocha, "On the implementation of the probabilistic logic programming language problog," *Theory and Practice of Logic Programming*, vol. 11, no. 2-3, pp. 235–262, 2011.

[22] L. De Raedt and K. Kersting, "Probabilistic logic learning," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 31–48, 2003.

[23] D. Lavanya and K. U. Rani, "Ensemble decision tree classifier for breast cancer data," *International Journal of Information Technology Convergence and Services*, vol. 2, no. 1, p. 17, 2012.

[24] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[25] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type ii diabetes," in *2011 International conference on innovations in information technology*, pp. 303–307, IEEE, 2011.

[26] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.