5-2013

A Knowledge-based Clinical Toxicology Consultant for Diagnosing Multiple Exposures

Joel D. Schipper
Embry-Riddle Aeronautical University, schippej@erau.edu

Douglas D. Dankel II
University of Florida

A. Antonio Arroyo
University of Florida

Jay L. Schauben
Florida Poison Information Center - Jacksonville, Shands Jacksonville Medical Center/University of Florida Health Science Center - Jacksonville

Scholarly Commons Citation

Schipper, J. D., Dankel II, D. D., Arroyo, A. A., & Schauben, J. L. (2013). A Knowledge-based Clinical Toxicology Consultant for Diagnosing Multiple Exposures. *Artificial Intelligence in Medicine, 58*(1). https://doi.org/10.1016/j.artmed.2013.02.002

Published in *Artificial Intelligence in Medicine*, Volume 58, Issue 1, May 2013, Pages 15-21 (http://dx.doi.org/10.1016/j.artmed.2013.02.002).

A knowledge-based clinical toxicology consultant for diagnosing multiple exposures

Joel D. Schipper[a,*], Douglas D. Dankel II[b], A. Antonio Arroyo[c], Jay L. Schauben[d]

[a] Electrical and Computer Engineering, Embry-Riddle Aeronautical University, 3700 Willow Creek Road, Prescott, AZ 86301, USA

[b] Computer and Information Science and Engineering, Box 116120, University of Florida, Gainesville, FL 32611, USA

[c] Electrical and Computer Engineering, Box 116200, University of Florida, Gainesville, FL 32611, USA

[d] Florida Poison Information Center – Jacksonville, Shands Jacksonville Medical Center/University of Florida Health Science Center – Jacksonville, 655 West 8th Street, Box C23, Jacksonville, FL 32209, USA

[*] Corresponding author. Tel.: +1 928 777 4205; Fax: +1 928 777 6945. *E-mail address*: joel.schipper@erau.edu (J. Schipper).

Correspondence Address:
E-mail: joel.schipper@erau.edu

Joel Schipper
Electrical & Computer Engineering
3700 Willow Creek Road
Prescott, AZ 86301
USA

Abstract

Objective: This paper presents continued research toward the development of a knowledge-based system for the diagnosis of human toxic exposures. In particular, this research focuses on the challenging task of diagnosing exposures to multiple toxins. Although only 10% of toxic exposures in the United States involve multiple toxins, multiple exposures account for more than half of all toxin-related fatalities. Using simple medical mathematics, we seek to produce a practical decision support system capable of supplying useful information to aid in the diagnosis of complex cases involving multiple unknown substances.

Methods: The system is automatically trained using data mining techniques to extract prior probabilities and likelihood ratios from a database managed by the Florida Poison Information Center (FPIC). When supplied with observed clinical effects, the system produces a ranked list of the most plausible toxic exposures. During testing, the system diagnosed toxins at three levels: identifying the substance, identifying the toxin's major and minor categories, and identifying the toxin's major category alone. To enable comparison between these three levels, accuracy was calculated as the percentage of exposures correctly identified in top 10% of trained diagnoses.

Results: System evaluation utilized a dataset of 8,901 multiple exposure cases and 37,617 single exposure cases. Initial system testing using only multiple exposure cases yielded poor results, with diagnosis accuracies ranging from 18.5-50.1%. Further investigation revealed that the system's inability to diagnose multiple disorders resulted from insufficient data and that the clinical effects observed in multiple exposures are dominated by a single substance. Including single exposures when training, the system achieved accuracies as high as 83.5% when

diagnosing the primary contributors in multiple exposure cases by substance, 86.9% when diagnosing by major and minor categories, and 79.9% when diagnosing by major category alone.

Conclusions: Although the system failed to completely diagnose exposures to multiple toxins, the ability to identify the primary contributor in such cases may prove valuable in aiding medical personnel as they seek to diagnose and treat patients. As time passes and more cases are added to the FPIC database, we believe system accuracy will continue to improve, producing a viable decision support system for clinical toxicology.

KEYWORDS
Knowledge-based systems;
Decision support systems;
Data mining;
Multiple disorder diagnosis;
Differential diagnosis;
Clinical toxicology

1. Introduction

This paper discusses ongoing research in developing a knowledge-based system to serve as a decision support system for clinical toxicology. The system is automatically trained using data mining techniques to extract likelihood ratios and prior probabilities from a database supplied by the Florida Poison Information Center (FPIC). After training, the user enters the clinical effects (i.e., signs and symptoms) observed in a patient and the system returns a differential diagnosis of plausible toxic exposures in the form of a ranked list. A brief overview of the system is given in [1,2], while [3] offers a detailed description of system functionality when diagnosing exposures to a single toxin. The research presented here expands on [3] by exploring the diagnosis of

multiple toxic exposures. The results reveal intriguing insights into the diagnosis of multiple exposures in the field of toxicology.

1.1 Motivation for diagnosing multiple disorders

Poison control centers offer free consultations with toxicologists and other specialists in the field of toxicology. In many cases, consultations are a simple matter, consisting mainly of matching clinical effects that are known to be directly associated with the mechanisms and behaviors of one class of toxin. Cases that toxicologists find difficult tend to consist of multiple unknown toxins interacting to produce clinical effects that cannot be matched with any single substance. If all substances interacted linearly, determining multiple unknown drugs by their clinical effects would amount to identifying the drug combinations that, when summed together, produce the observed results. Unfortunately, many drug interactions are non-linear, interacting synergistically or antagonistically. Some drug combinations cause a dramatic increase in symptom severity, some mask symptoms normally observed with one of the drugs, and some can cause symptoms that normally would not appear with any of the drugs individually. Little documentation exists for the majority of toxic exposure combinations that can occur and, although many established methods for designing knowledge-based systems exist (e.g., rule-based systems, case-based reasoning, etc.), none have fully solved the problem of diagnosing multiple disorders. Additionally, only a limited number of knowledge-based systems exist in the field of toxicology, the most prominent being a French system called SETH [4,5], a Bulgarian system called MEDICOTOX-CONSILIUM [6], and an Inreca (Induction and Reasoning from Cases) system focused on Russian intoxications [7]. None of these publications thoroughly discuss the diagnosis of multiple disorders, nor are the systems readily available for use by American toxicologists.

Beyond the motivation of developing technology to address an unsolved diagnostic problem, the more important concern is saving lives. The 2010 National Poison Data System (NPDS) report shows that although only 10% of reported human exposures involved multiple toxins, multiple toxins accounted for 58.6% of all exposure-related fatalities [8]. Being able to address multiple exposures is an important concern for the preservation of human lives. A knowledge-based system can aid in addressing multiple exposures by effectively making the relevant information in poison control center databases available to the toxicologist. The goal of the knowledge-based system presented in this paper is not to replace the toxicologist, but to act as a powerful consulting tool providing case-based summary data for all medical personnel that may encounter toxic exposure cases. Human beings have senses and intuition that are important for diagnosis, which computers cannot replicate. However, by offering speculative advice, the system may facilitate more accurate and timely diagnoses.

1.2 Approaches to diagnosing multiple disorders

Four primary approaches have been used when developing knowledge-based systems for multiple disorder diagnosis: Bayesian methods, case-based reasoning, set covering, and diagnostic scores. Note that, although we divide these systems into four types for the sake of discussion, many systems may contain aspects from multiple approaches.

Bayesian methods revolve around Bayes' rule of conditional probability, which requires statistical independence of the clinical effects used for diagnosis. As a result, much research has focused on the generalization of Bayes' rule to account for dependencies within a domain (e.g., [9,10]). Of note, Bayesian belief networks were developed to enable dependencies to be included in a system's probability calculations [11]. Examples of systems using belief networks to

diagnose multiple disorders include HEPAR II [12,13], MUNIN [14], and a system developed in the Netherlands [15].

Case-based reasoning enables systems to effectively create themselves from historical cases, unlike most complex models (including Bayesian belief networks) that generally require knowledge acquisition from experts [16]. Since case-based reasoning is an approach to system development rather than a method for reconciling uncertainty and probabilistic dependencies, many case-based reasoning systems also make use of other methods. Examples of systems using case-based reasoning to diagnose multiple exposures include ADAPtER [17] and research based on the SONOCONSULT knowledge base [16,18,19].

Set covering is a method that seeks to find combinations of disorders that can account for observed clinical effects. The simplicity and elegance of the approach makes it one of the most promising areas in research relating to multiple disorder diagnosis. Research using set covering to diagnose multiple exposures has been performed by Reggia et al. [20], Peng and Reggia [21-23], Wu [24,25], and others [26,27].

Finally, diagnostic scores are commonly used in the medical field when diagnosing multiple disorders. When forming a diagnosis, the physician gathers a list of all the clinical effects observed in the patient. Each clinical effect has a score based on its correlation to a specific disorder. Calculating the sum of these scores yields a final diagnostic score, which corresponds to the patient's risk of having a specific disorder. Research based on SONOCONSULT [28,29] provides an excellent example of a system using diagnostic scores to diagnose multiple disorders. This case-based system semi-automatically (i.e., the system generates its rules automatically but still requires an expert to oversee its development and adjust parameters as necessary to ensure the system functions properly) learns diagnostic rules for the

field of sonography. Since "understandability and interpretability of…learned models is of prime importance," the system utilizes diagnostic scores because, "ideally, [a] learning method constructs knowledge in the same representation the human expert favors" [28].

2. System approach

Our system was developed in Microsoft Access 2002 and programmed using Visual Basic and SQL. The system is a hybrid, merging concepts from three of the approaches discussed above. Like case-based reasoning, the system relies entirely on a database of cases for diagnosis. Like Bayesian approaches, the core of the system is based on conditional probability calculations (i.e., likelihood ratios). (In fact, replacing current system calculations with the odds-ratio form of Bayes' theorem [30] does not affect the order of the resultant differential diagnosis.) Like diagnostic scores, the system seeks to use established representations recognized in the field of medicine (i.e., differential diagnoses, likelihood ratios, and pre-test probabilities).

Likelihood ratios and pre-test probabilities (prior probabilities) are commonly used throughout the medical field and readily understood by healthcare professionals. Furthermore, by assuming that each clinical effect functions as an independent diagnostic test, the calculations are easily combined via multiplication. For these reasons, the system utilizes tables of likelihood ratios and prior probabilities to produce its differential diagnosis. The basic likelihood ratio, LR^+, is calculated as:

$$LR^+ = \frac{\left(\frac{TP}{TP + FN}\right)}{\left(\frac{FP}{FP + TN}\right)}, \tag{1}$$

where *TP* represents true positive, *TN* represents true negative, *FP* represents false positive, and *FN* represents false negative test results [30]. To prevent multiply-by-zero and divide-by-zero errors, a pseudocount [31] is added to the likelihood ratio:

$$LR^{+}_{Adj} = \frac{\left(\frac{TP + \Delta}{TP + \Delta + FN + \Delta}\right)}{\left(\frac{FP + \Delta}{FP + \Delta + TN + \Delta}\right)}, \quad (2)$$

where Δ is a small, positive constant. A Δ of 0.01 is used, unless otherwise noted. The system's primary resource when generating a differential diagnosis is an exhaustive table of likelihood ratios, calculated using (2), relating every individual clinical effect to every possible toxic exposure diagnosis. In addition, the system uses a table containing the prior probability, *P*, of each toxin, which is calculated as:

$$P = \frac{Cases}{Total}, \quad (3)$$

where *Cases* is the number of cases involving a particular substance and *Total* is the total number of exposure cases in the database [32].

The system's likelihood ratios and prior probabilities were automatically trained on five years of case data supplied by the FPIC in Jacksonville. After cleaning the database using data mining techniques, 37,617 single exposure cases and 8,901 multiple exposure cases remained for system training. The cases in the FPIC database conform to the national standards set forth by the National Poison Data System (NPDS), guaranteeing that every exposure record includes the clinical effects and final diagnosis associated with the case. These standards not only accommodate the training and testing of the system, they also facilitate the portability of the system to other poison control centers around the United States.

NPDS categorizes each substance as belonging to a major and minor category. For example, black widow spider poison is part of the spider bites minor category, which is part of the bites and envenomations major category. Even if a specific substance cannot be determined, identifying a toxin's general category can aid in the prompt treatment of a patient. Because of this, the system was trained to diagnose at three levels: substance, major and minor categories, and major category alone. Furthermore, the system was tested at three levels of medical outcomes: exposures of minor severity or worse, moderate severity or worse, and major severity or worse.

System testing is performed using 10-fold cross-validation. Based on the observed clinical effects, combined likelihood ratios (including prior probabilities) are calculated by multiplication and a differential diagnosis generated in the form of a ranked list. To enable comparison between different diagnostic levels during testing, accuracy is calculated as the percentage of cases identified correctly in the top 10% of the trained diagnoses. Unless otherwise noted, trained diagnoses are limited to include only exposures for which a minimum of 10 recorded cases appeared in the database. For a more detailed discussion of system development and design principles, see [3].

3. System testing and results

3.1 Diagnosing multiple exposures using solely multiple exposure cases

During the initial phase of testing, all multiple exposure cases were extracted from the database. NPDS standards require that each substance involved in a toxic exposure be assigned a sequence number that ranks the substance in accordance with its relative contribution to the observed clinical effects. To prevent combinatorial explosion in the initial attempts to diagnose multiple disorders, only the primary and secondary contributors in each multiple exposure case

were considered. NPDS standards also require that substances be recorded by a product specific code as well as a generic substance code. From this requirement, a problem arises. When determining the number of substances involved in an exposure, the FPIC database uses the product specific code. As a result, two products marketed by different companies are listed as separate substances, even if their active ingredient is the same. When cleaning the data, if the generic substance codes for the top three contributing substances were identical, the case was removed from the dataset. If the first two generic substance codes were identical but the third was different, the third substance was treated as the secondary contributor for the case. Finally, the multiple exposure cases were filtered so that only cases followed to a known outcome that produced at least minor effects in the patient were used to train and test the system. The resulting cleaned dataset contains 8,901 multiple exposure cases.

When generating the multiple exposure system, each pair of primary and secondary contributors was trained as a single diagnosis. The original results from training and testing the system on multiple exposure cases are displayed in the first column of Table 1. With an accuracy ranging from 28.3-50.1%, the system's deplorable performance is painfully obvious. To further explore this failure, the system was tested by altering the cutoff for the minimum number of recorded cases required to train a particular diagnosis from 10 to 15, 20, and 25. The results of these tests show a similar lack of accuracy (Table 1). Looking at the rows in the table from left to right, we can see that the performance gradually decays as the cutoff value increases. As discussed in [3], such an observation is expected because an increase in the minimum number of exposure cases lowers the number of diagnoses included in the top 10%. The most interesting characteristic of the data in Table 1 is that as the severity increases, the accuracy decreases. This observation is contrary to the results observed in the single exposure system [3]. Normally,

system accuracy increases with severity because more severe cases contain more clinical effects, making diagnosis easier for the system.

There are a number of plausible explanations for why accuracy might decrease with severity, but two are particularly compelling. The first explanation is that the decrease in accuracy is caused by the non-linear interactions between multiple toxins. As the severity of an exposure increases, there is greater opportunity for a combination of toxins to produce effects not normally associated with any of the toxins individually. This could lower the accuracy of the system because the clinical effects would behave more erratically and might not correspond to the majority of cases. The second explanation is that the decrease in accuracy is simply caused by lack of quality data. As the severity cutoff becomes more stringent, fewer cases are tested against the system, leading to a poor sampling and quite possibly lower accuracies on average. Lack of quality data could account for both the low accuracy observed overall as well as the decrease in accuracy as the severity increases.

Another parameter that might contribute to the system's poor accuracy is the pseudocount (Δ) introduced in (2). The pseudocount is meant primarily to safeguard against multiply-by-zero and divide-by-zero errors, however, a small training set might cause Δ to adversely influence the diagnostic results. Table 2 compares the original system accuracy, when using a Δ of 0.01, to accuracies calculated with a Δ of 0.1 and 0.001. It was discovered that increasing Δ to 0.1 causes an average decrease in accuracy of 1.6%, while decreasing Δ to 0.001 causes an average increase in accuracy of only 0.1%. These results imply that a Δ of 0.01 yields satisfactory relative performance compared to other Δ parameters that might be selected.

In an attempt to improve accuracy and better understand the system's poor performance, a number of system variations were tested. The resulting accuracies for these systems are presented in Table 3, where the column labeled "original accuracies" represents the original system. The first column of accuracies displays the results for a system that assumes all trained diagnoses are equally likely by omitting prior probabilities from its calculations. As expected, the system performs worse than the original. However, the results of this test do reveal a few important insights. Note that, unlike the original, the accuracies for diagnosis by substance as well as major and minor categories increase as severity increases. The significance of this observation is that the system is indeed processing clinical effects correctly. Thus, the accuracies decreasing with increased severities in the original testing are not due to the non-linear interactions of multiple substances. Rather, the results imply that the prior probability is dominating the original diagnoses. The most likely cause for this problem is lack of quality data. Additionally, the fact that diagnosis by major category alone still displays a decreasing accuracy with increasing severity fits the explanation. Major categories cover a broad variety of substances, making it difficult to train a general model that properly fits the major category as a whole. The problem is compounded when attempting to identify two different major categories in the same diagnosis.

Another problem that could contribute to the low accuracy of the system is that multiple exposure cases can consist of more than two substances. Since the system only considers the primary and secondary contributors, any additional substances involved could affect the clinical effects in a manner not normally predicted in a case only involving two substances. To improve the quality of the training data, a system was created based solely on cases where exactly two substances are involved. The system accuracy is reported in Table 3 under the column titled

"double exposures." Although this approach improves data quality, it also reduces the amount of training cases from 8,901 to 5,149, a data reduction of over 40%. The end results yield a nominal increase in the average accuracy of only 0.7%.

Further attempts to improve accuracy resulted in two more variations of the system. The original system requires the correct identification of both primary and secondary contributors for a diagnosis to be considered successful. The first variation relaxes the constraints of the original system by allowing the order of the primary and secondary contributing substances to be reversed. Thus, diagnosing a test case with a primary contributor of A and a secondary contributor of B as having a primary contributor of B and a secondary contributor of A is considered an accurate diagnosis. As seen in Table 3 under the column labeled "order reversed," the relaxed diagnosis criteria increase accuracy by an average of 8.9%. Unfortunately, the resulting system is still not viable, having only achieved a maximum accuracy of 56.0%. The second variation on the original system attempted to improve accuracy by allowing the system to count any diagnosis as a correct match if the primary contributor matched the primary contributor of the test case, regardless of the secondary contributors involved. As shown in Table 3 under the column labeled "primary correct," this increases the system's accuracy drastically, yielding a maximum accuracy of 82.9%. It should be noted that these results are falsely optimistic because the most common substances involved in multiple exposures are the primary contributors for many different substance combinations. As a result, a number of different possible diagnoses could be considered "correct" diagnoses for any single test case. Additionally, diagnosing multiple exposures by substance has a maximum accuracy of 64.8%, which is not an outstanding number. In spite of these shortcomings, the final system test seems to indicate that the primary contributor might be the dominating force in most multiple exposure cases. For that

reason, the research presented in the following section focuses on diagnosing the primary contributor.

3.2 Diagnosing multiple exposures with single exposure cases

The findings in the previous section seem to indicate that the clinical effects observed in most multiple exposure cases are dominated by the clinical effects associated with the primary contributor. To test this hypothesis, a system trained entirely on single exposures was examined to see if it could accurately diagnose the primary contributor in multiple exposure cases. The first column of Table 4 shows the accuracy of the system when diagnosing the primary contributor for every multiple exposure case. The next column shows the results when the test cases are limited to double exposures. With accuracies reaching as high as 84.9%, the results confirm that the clinical effects observed in most multiple exposure cases are indeed dominated by those associated with the primary contributor. Furthermore, the evidence indicates that the poor performance observed in the system trained solely on multiple exposure cases was not due to non-linear interactions between multiple toxins. As discussed in the previous section, the remaining explanation for the system failure is lack of sufficient data.

To test whether lack of data caused the poor performance observed in the system trained solely on multiple exposures, a system was trained using a combination of multiple exposures and single exposures to diagnose the primary contributor in multiple exposure cases. For training purposes, each multiple exposure was treated as a single exposure case with the primary contributor as the correct diagnosis. The system was then tested using 10-fold cross-validation on the multiple exposure dataset with the modification that all single exposure cases were included in the training set for every iteration. In a similar manner, a system trained on a combination of double exposures and single exposures was tested to see if it could identify the primary

contributor in double exposure cases. The results of these two tests are displayed in the last two columns of Table 4. On average, the accuracy increased by 3.3% when diagnosing multiple exposures and 1.8% when diagnosing only double exposures. These results indicate that valuable information capable of yielding greater than 80% accuracy is contained in the multiple exposure cases. Moreover, these results are consistent with the explanation that the system failure when training on multiple exposures alone was due to lack of sufficient data. It is also interesting to note that the system performed slightly better diagnosing multiple exposures, which generally should contain more extraneous clinical effects, than when diagnosing double exposures. The explanation is that training with multiple exposures included the information from approximately 8,011 cases per diagnosis cycle whereas training with double exposures included approximately 4,634 cases per diagnosis cycle. Presumably, having the same number of double exposures as multiple exposures would result in the double exposures performing better. A similar observation can be made of the data presented in Table 5.

The first two columns in Table 5 display the accuracies of a system trained solely on single exposure cases and tested against the secondary contributor for both multiple and double disorder cases. With average accuracies of 69.1% and 66.3%, the system performance is not stellar, however, it is high enough to raise a question: If the clinical effects in multiple exposure cases are dominated by the primary contributor, why is the accuracy in diagnosing the secondary contributor so high? Recall that during data cleaning all multiple exposure cases involving only products with the same generic substance code are removed from the dataset. This cleaning is only performed at the substance level. It is still likely that many multiple exposure cases consist of primary and secondary substances that share the same major and minor categories. Belonging to the same category makes it much more likely that the two substances exhibit similar clinical

effects. Examining the data, it was determined that 21.0% of the primary and secondary contributors in all multiple exposure cases belonged to the same major category and 11.6% belonged to the same minor category as well. Likewise, 21.9% of all primary contributors in double exposure cases belonged to the same major category and 11.1% belonged to the same minor category. Because these cases are more likely to be diagnosed correctly based on the primary contributor, the accuracies are falsely optimistic.

The last two columns in Table 5 show the accuracies of a system trained on a combination of single exposures and the secondary contributors for either multiple exposures or double exposures. The addition of the secondary contributors improves the average system accuracy by 9.1% for multiple exposure diagnosis and 9.3% for double exposure diagnosis. Such a significant jump in accuracy attests that, although dominated by the primary contributor's clinical effects, secondary contributors do produce enough clinical effects that the system can be trained to at least recognize the most common multiple exposure combinations. Although some of the accuracy can be accounted for by prior probabilities, the results give hope that further research might enable reasonably accurate identification of secondary contributors.

The final step necessary to fully explore the impact of combining multiple exposure cases with single exposure cases was to train a system with the combined data and use it to diagnose only single exposure cases (Table 6). The first column shows the accuracy of a system trained on single exposures alone when diagnosing single exposures. The second and third columns display the accuracies for systems trained on single exposures along with the primary contributors for either multiple or double exposures. The last two columns contain the accuracies of systems trained on single exposures along with the secondary contributors for either multiple or double exposures. Interestingly, those systems trained with the primary contributors increased the

average system accuracy from 74.6% to 74.9% when including multiple exposures and 75.1% when including double exposures. Although a minor increase, it is an increase nonetheless and lends further support to the conclusion that the clinical effects in multiple exposure cases are dominated by the primary contributor. Furthermore, the average accuracy for systems trained with secondary contributors decreased from 74.6% to 74.2% when including multiple exposures and 74.4% when including double exposures. A lower accuracy is to be expected since training on the secondary contributor would associate clinical effects caused by the primary contributor with the secondary contributor instead. The minimal change in accuracy can be partially explained by the multiple and double exposures that involve closely related substances from the same major and minor categories, as discussed above. Additionally, on average 33,855 single exposure cases were used to train the system on each cycle. The added 8,901 multiple exposure cases or 5,149 double exposure cases only account for approximately 20.8% and 13.2% of the training cases.

4. Conclusion

The research presented continues the development of a prototype knowledge-based system by investigating the diagnosis of exposures to multiple toxins. The system intentionally uses simple computations, following the philosophy that "simpler, even trivial, processes are better than complicated ones if they are enough for the job of discovery" [33]. Such simplicity will become necessary for scalability as the FPIC database grows in size. Although lack of multiple exposure data inhibited the diagnosis of more than one substance at a time, system testing revealed that the clinical effects observed in multiple exposures tend to be dominated by a single substance, called the primary contributor. Training the system on a combined training set of both single exposures and primary contributors from multiple exposure cases yielded performances as

high as 86.9% accuracy when diagnosing primary contributors. More specifically, 86.9% of the cases were diagnosed in the top 13 out of 129 possible major and minor category combinations. Being able to diagnose the disorder causing the most detrimental clinical effects is certainly valuable. Once the primary contributor is treated, it becomes easier to identify the other contributors in a multiple exposure case. Furthermore, there is hope that the accuracy when simultaneously diagnosing multiple exposures will improve with the collection of more data.

5. Future work

From the outset, a major objective of the research was to bypass the knowledge acquisition bottleneck by generating a knowledge-based system capable of producing meaningful and useful results without the need for an active, overseeing expert. This design principle inherently limited the designer from making any changes that required even a fundamental knowledge of toxicology. Now that the prototype is complete, several changes can be implemented for the betterment of the system. First, useless substance diagnoses, such as the "unknown drug" diagnosis, should be removed. Second, redundant substances, such as "aspirin: pediatric formulation," "aspirin: unknown if adult or pediatric formulation," and "aspirin: adult formulation," should be consolidated into a single diagnosis. Third, the category divisions could be examined by a toxicologist to create groupings based primarily on clinical effects. For example, most opioids tend to exhibit similar clinical effects whereas the effects associated with spider bites vary greatly depending on the species of spider. Intelligently restructuring diagnostic groupings could greatly increase the accuracy and utility of the knowledge-based consultant.

After refining the system, the next step is to field test the system at the FPIC by implementing the system directly on the FPIC's dedicated SQL server. Based on these results, further improvements can be implemented. One possible concern is that, although the system

may perform well on toxic exposure cases as a whole, it may be more beneficial for the system to specialize on more difficult and deadly problems. In other words, it may be better to sacrifice accuracy on simple, routine exposures to increase the accuracy of the system on exposures that are dangerous and difficult to diagnose.

Finally, the system could be converted into a program for knowledge discovery within toxicology. When training on cases in the database, the system identifies relationships between specific exposures and their clinical effects. While many of these relationships are already known, it is quite possible that the system is discovering new relationships that were previously undocumented. This is particularly true when characterizing multiple exposure cases, many of which have little documentation. Examining the relationships within a trained system could lead to new discoveries in the field of toxicology, as has been done in other medical fields [34,35].

6. Acknowledgments

The authors thank Dausear "Dar" McRae and the FPIC in Jacksonville for their willingness to provide data, technical support, and consultation.

References

[1] J. D. Schipper, J. L. Schauben, D. D. Dankel, A. A. Arroyo, and D. R. Sollee, Differential toxicological diagnoses using a computerized knowledge-based model, *Clinical Toxicology 46* (2008) 612.

[2] J. D. Schipper, J. L. Schauben, D. D. Dankel, A. A. Arroyo, and D. R. Sollee, A knowledge-based consultant for human toxic exposures, *Clinical Toxicology 47* (2009) 465-466.

[3] J. D. Schipper, D. D. Dankel, A. A. Arroyo, and J. L. Schauben, A knowledge-based clinical toxicology consultant for diagnosing single exposures, *Artificial Intelligence in Medicine 55* (2012) 87-95.

[4] S. Darmoni, P. Massari, J. Droy, N. Mahe, T. Blanc, E. Moirot, et al., SETH: an expert system for the management on acute drug poisoning in adults, *Computer Methods and Programs in Biomedicine 43* (1994) 171-176.

[5] S. Darmoni, P. Massari, J. Droy, T. Blanc, and J. Leroy, Functional evaluation of Seth: an expert system in clinical toxicology, in: P. Barahona, M. Stefanelli, and J. Wyatt, eds., *Artificial intelligence in medicine: 5th conference on artificial intelligence in medicine Europe* (Springer, Berlin, 1995) 231-238.

[6] A. Monov, I. Iordanova, P. Zagorchev, V. Vassilev, M. Nissimov, R. Kojuharov, et al., MEDICOTOX CONSILIUM - an expert system in clinical toxicology, in: K. Lun, P. Degoulet, T. Piemme, and O. Rienhoff, eds., *Proceedings of the 7th world congress on medical informatics* (North-Holland Publishing Company, Amsterdam, 1992) 610-614.

[7] K. Althoff, R. Bergmann, S. Wess, M. Manago, E. Auriol, O. Larichev, et al., Case-based reasoning for medical decision support tasks: The Inreca approach, *Artificial Intelligence in Medicine 12* (1998) 25-41.

[8] A. Bronstein, D. Spyker, L. Cantilena, J. Green, B. Rumack, and R. Dart, 2010 annual report of the American Association of Poison Control Centers' National Poison Data System (NPDS): 28th annual report, *Clinical Toxicology 49* (2011) 910-941.

[9] M. Ben-Bassat, R. Carlson, V. Puri, M. Davenport, J. Schriver, M. Latif, et al., Pattern-based interactive diagnosis of multiple disorders: the MEDAS system, *IEEE Transactions on Pattern Analysis and Machine Intelligence 2* (1980) 148-160.

[10] M. Ben-Bassat, D. Campell, A. MacNeil, and M. Weil, Evaluating multimembership classifiers: a methodology and application to the MEDAS diagnostic system, *IEEE Transactions on Pattern Analysis and Machine Intelligence 5* (1983) 225-229.

[11] J. Pearl, Bayesian networks: A model of self-activated memory for evidential reasoning, UCLA Technical Report CSD-850017, Los Angeles, CA, 1985.

[12] A. Onisko, M. Druzdzel, and H. Wasyluk, Extension of the HEPAR II model to multiple-disorder diagnosis, in: M. Klopotek, M. Michalewicz, and S. Wierzchon, eds., *Intelligent information systems: proceedings of the IIS'2000 symposium* (Physica-Verlag, Heidelberg, 2000) 303-313.

[13] A. Onisko, M. Druzdzel, and H. Wasyluk, Learning Bayesian network parameters from small data sets: application of Noisy-OR gates, *International Journal of Approximate Reasoning 27* (2001) 165-182.

[14] M. Suojanen, S. Andreassen, and K. Olesen, A method for diagnosing multiple diseases in MUNIN, *IEEE Transactions on Biomedical Engineering 48* (2001) 522-532.

[15] L. van der Gaag and M. Wessels, Efficient multiple-disorder diagnosis by strategic focusing, Utrecht University Technical Report UU-CS-1994-23, Utrecht, The Netherlands, 1994.

[16] M. Atzmueller, J. Baumeister, F. Puppe, W. Shi, and J. Barnden, Case-based approaches for diagnosing multiple disorders, in: V. Barr and Z. Markov, eds., *Proceedings of the 17th international Florida artificial intelligence research society conference* (AAAI Press, Menlo Park, CA, 2004) 154-159.

[17] L. Portinale and P. Torasso, ADAPtER: an integrated diagnostic system combining case-based and abductive reasoning, in: M. Veloso and A. Aamodt, eds., *Proceedings of the 1st international conference on case-based reasoning research and development* (Springer-Verlag, London, UK, 1995) 277-288.

[18] J. Baumeister, M. Atzmueller, and F. Puppe, Inductive learning for case-based diagnosis with multiple faults, in: S. Craw and A. Preece, eds., *Advances in case-based reasoning: 6th European conference* (Springer, Berlin, 2002) 28-42.

[19] M. Atzmueller, J. Baumeister, and F. Puppe, Evaluation of two strategies for case-based diagnosis handling multiple faults, in: M. Nick and K.-D. Althoff, eds., *Proceedings of the 2nd German workshop on experience management* (CEUR Workshop Proceedings, Aachen, Germany, 2003).

[20] J. Reggia, D. Nau, and P. Wang, Diagnostic expert systems based on a set covering model, *International Journal of Man-Machine Studies 19* (1983) 437-460.

[21] Y. Peng and J. Reggia, Plausibility of diagnostic hypotheses: the nature of simplicity, in: T. Kehler and S. Rosenschein, chairmen, *Proceedings of the 5th national conference on artificial intelligence* (AAAI Press, Menlo Park, CA, 1986) 140-145.

[22] Y. Peng and J. Reggia, A probabilistic causal model for diagnostic problem solving - part I: integrating symbolic causal inference with numeric probabilistic inference, *IEEE Transactions on Systems, Man, and Cybernetics 17* (1987) 146-162.

[23] Y. Peng and J. Reggia, A comfort measure for diagnostic problem solving, *Information Sciences 47* (1989) 149-184.

[24] T. Wu, A problem decomposition method for efficient diagnosis and interpretation of multiple disorders, *Computer Methods and Programs in Biomedicine 35* (1991) 239-250.

[25] T. Wu, Efficient diagnosis of multiple disorders based on a symptom clustering approach, in: T. Dietterich and W. Swartout, chairmen, *Proceedings of the 8th national conference on artificial intelligence* (AAAI Press, Menlo Park, CA ,1990) 357-364.

[26] S. Vinterbo and L. Ohno-Machado, A genetic algorithm approach to multi-disorder diagnosis, *Artificial Intelligence in Medicine 18* (2000) 117-132.

[27] J. Baumeister, D. Seipel, and F. Puppe, Incremental development of diagnostic set-covering models with therapy effects, in: G. Kern-Isberner, T. Lukasiewicz, and E. Weydert, eds., *Proceedings of the KI-2001 workshop on uncertainty in artificial intelligence* (Linkping University Electronic Press, Linkoping, Sweden, 2001).

[28] M. Atzmueller, J. Baumeister, and F. Puppe, Inductive learning of simple diagnostic scores, in: P. Perner, R. Brause, H.-G. Holzhutter, eds., *Medical data analysis: 4th international symposium* (Springer, Berlin, 2003) 23-30.

[29] M. Atzmueller, J. Baumeister, and F. Puppe, Quality measures for semi-automatic learning of simple diagnostic rule bases, in: D. Seipel, M. Hanus, U. Geske, and O. Bartenstein, eds., *15th international conference on applications of declarative programming and knowledge management* (Springer, Berlin, 2004) 65-78.

[30] D. Owens and H. Sox, Medical decision-making: probabilistic medical reasoning, in: E. Shortliffe and L. Perreault, eds., *Medical informatics: computer applications in health care and biomedicine* (Springer-Verlag, New York, NY 2001) 76-131.

[31] R. Maxion and T. Townsend, Masquerade Detection Using Truncated Command Lines, in: *Proceedings of the international conference on dependable systems and networks* (IEEE Computer Society, Los Alamitos, CA, 2002) 219-228.

[32] R. Duda, P. Hart, and D. Stork, Bayesian decision theory, *Pattern Classification*, 2nd ed. (John Wiley & Sons, Inc., New York, NY, 2001).

[33] R. Valdes-Perez, Principles of human-computer collaboration for knowledge discovery in science, *Artificial Intelligence 107* (1999) 335-346.

[34] S. Brossette, A. Sprague, J. Hardin, K. Waites, W. Jones, and S. Moser, Association rules and data mining in hospital infection control and public health surveillance, *Journal of the American Medical Informatics Association 5* (1998) 373-381.

[35] J. Breault, C. Goodall, and P. Fos, Data mining a diabetic data warehouse, *Artificial Intelligence in Medicine 26* (2002) 37-54.

Table 1

Accuracy (varying cutoff for training) of system trained and tested on multiple exposures

		Minimum exposure cases			
Diagnosed by	Severity	10	15	20	25
Substance	Minor	33.5%	30.4%	29.0%	27.6%
	Moderate	30.0%	26.9%	25.3%	22.9%
	Major	28.3%	23.3%	21.8%	18.5%
Major and minor categories	Minor	47.3%	43.6%	39.5%	38.2%
	Moderate	45.9%	42.1%	37.6%	36.5%
	Major	37.6%	34.5%	30.9%	30.6%
Major category	Minor	50.1%	46.8%	45.7%	43.4%
	Moderate	47.2%	44.1%	43.0%	40.4%
	Major	43.0%	39.6%	38.2%	36.5%
	Average	40.3%	36.8%	34.5%	32.7%

Table 2

Accuracy (varying Δ) of system trained and tested on multiple exposures

Diagnosed by	Severity	$\Delta = 0.1$	$\Delta = 0.01$	$\Delta = 0.001$
Substance	Minor	32.3%	33.5%	33.5%
	Moderate	29.0%	30.0%	29.8%
	Major	26.4%	28.3%	28.0%
Major and minor categories	Minor	46.5%	47.3%	47.4%
	Moderate	44.6%	45.9%	46.1%
	Major	35.0%	37.6%	38.3%
Major category	Minor	49.4%	50.1%	50.2%
	Moderate	46.1%	47.2%	47.2%
	Major	39.5%	43.0%	43.4%
	Average	38.8%	40.3%	40.4%

Table 3

Accuracy comparison of various systems for multiple exposure diagnosis

Diagnosed by	Exposure severity	No prior probability	Original accuracies	Double exposures	Order reversed	Primary correct
Substance	Minor	16.5%	33.5%	35.3%	42.4%	64.8%
	Moderate	17.5%	30.0%	30.9%	40.3%	63.0%
	Major	23.9%	28.3%	28.5%	39.8%	63.7%
Major and minor categories	Minor	21.7%	47.3%	47.1%	54.0%	82.7%
	Moderate	23.0%	45.9%	45.1%	53.7%	82.9%
	Major	23.5%	37.6%	42.3%	49.4%	81.2%
Major category	Minor	24.2%	50.1%	50.8%	56.0%	81.3%
	Moderate	23.7%	47.2%	47.1%	54.5%	81.5%
	Major	22.7%	43.0%	42.0%	53.3%	80.9%
	Average	21.9%	40.3%	41.0%	49.3%	75.8%

Table 4

Accuracy diagnosing primary contributors using single exposures

Diagnosed by	Severity	Singles diagnosing multiples	Singles diagnosing doubles	Combined diagnosing multiples	Combined diagnosing doubles
Substance	Minor	75.4%	75.2%	79.1%	77.5%
	Moderate	77.2%	77.3%	81.1%	79.3%
	Major	78.7%	81.8%	83.5%	83.1%
Major and minor categories	Minor	77.8%	76.4%	80.4%	78.2%
	Moderate	81.3%	80.4%	83.3%	81.6%
	Major	84.9%	84.9%	86.9%	86.2%
Major category	Minor	74.4%	74.9%	77.7%	76.9%
	Moderate	75.5%	76.2%	78.7%	78.5%
	Major	75.8%	78.3%	79.9%	80.5%
	Average	77.9%	78.4%	81.2%	80.2%

Table 5

Accuracy diagnosing secondary contributors using single exposures

Diagnosed by	Severity	Singles diagnosing multiples	Singles diagnosing doubles	Combined diagnosing multiples	Combined diagnosing doubles
Substance	Minor	69.6%	68.6%	77.6%	75.7%
	Moderate	70.5%	69.5%	79.7%	77.6%
	Major	69.5%	70.0%	81.6%	77.0%
Major and minor categories	Minor	67.8%	63.2%	78.3%	76.8%
	Moderate	73.0%	69.5%	82.4%	80.9%
	Major	77.6%	76.2%	86.2%	83.9%
Major category	Minor	62.1%	57.1%	71.4%	69.0%
	Moderate	64.4%	59.7%	72.9%	70.2%
	Major	67.4%	63.0%	74.3%	69.6%
	Average	69.1%	66.3%	78.2%	75.6%

Table 6

Comparison of system accuracies when diagnosing single exposure cases

Diagnosed by	Severity	Single exposures alone	Singles and multiples (primary)	Singles and doubles (primary)	Singles and multiples (secondary)	Singles and doubles (secondary)
Substance	Minor	68.3%	68.2%	68.4%	68.1%	68.2%
	Moderate	77.5%	78.2%	78.0%	77.4%	77.4%
	Major	80.7%	81.4%	81.4%	80.6%	80.8%
Major and minor categories	Minor	69.0%	68.9%	69.0%	68.6%	68.8%
	Moderate	77.6%	77.7%	78.0%	77.2%	77.5%
	Major	79.8%	80.6%	81.0%	80.6%	80.3%
Major category	Minor	68.8%	68.4%	68.9%	67.6%	67.9%
	Moderate	73.9%	74.3%	74.3%	73.4%	73.3%
	Major	76.2%	75.9%	76.8%	74.7%	75.0%
	Average	74.6%	74.9%	75.1%	74.2%	74.4%