Clinical Research in Pneumonia: Role of Artificial Intelligence

Timothy L Wiemken[1]*, Robert R Kelley[2], William A Mattingly[3] and Julio A Ramirez[3]

Introduction

Clinical research in pneumonia involves the creation and dissemination of new knowledge studying patients with pneumonia. The process of clinical research can be summarized in four steps: planning the study, performing of the study, analyzing the data, and disseminating study results. During the third step of data analysis, data are often examined to define if associations exist between independent variables (e.g. predictor variable or other variables in the model) and the dependent variable (e.g. outcome). This examination of the data is performed using two types of methods: 1) clinical analysis and 2) statistical analysis. During clinical analysis, the data are evaluated to define biological plausibility and clinical importance. During statistical analysis, the data are commonly evaluated to define statistical significance for the purposes of hypothesis testing, an approach termed 'frequentist statistics' [1].

Several limitations are recognized with the use of these traditional frequentist approaches in clinical research. The overemphasis on hypothesis testing, calculation of P-values as the primary means of significance assessment, and the convention of using the 5% level of statistical significance are all now considered limiting issues for defining causal effects; the actual impact of the independent variable (predictor) on the dependent variable (outcome). Here, the approach to analysis is based only on two mutually exclusive hypotheses, the null and alternative hypotheses, with results being either significant or nonsignificant. This leads to either rejecting the null hypothesis (significant P-value) or not rejecting the null hypothesis (non-significant P-value). Statistically significant P-values, those below an arbitrary cut point (often 0.05), can be obtained when the sample sizes are large but actual differences or associations are small (and potentially not clinically meaningful). On the other hand, non-significant results (e.g. P values over the cutoff) may be due to a small sample of patients (often termed lack of statistical power), or true non-significance. Therefore, when non-significant results are obtained using traditional hypothesis testing, it is not known if the results are due to the sample size (leading to reduced statistical power to detect the association or difference), or there truly is no association or difference between the predictor variable and the outcome.

Here, we propose adding machine learning, a branch of artificial intelligence, as a new methodology to analyze study results in pneumonia clinical research. Currently, few investigators use these methods as a replacement for traditional frequentist statistical methodologies.

Machine learning in small datasets

Until recently machine learning methods were considered useful only for 'big data' (e.g. millions of patients and hundreds of variables), and were used for purely predictive modeling (versus explanatory modeling). This led to most machine learning approaches being labeled as 'black boxes'. These black boxes were very accurate at predicting an outcome for a new individual after being trained on historical data, but were not able to provide an estimate of the impact of each variable on the outcome. This latter concept (explanatory modeling) has been a necessary part of clinical research for many decades. However, machine learning can now be used to estimate the impact of each variable on the outcome using a few novel methods. Newer algorithms are also much better at dealing with the smaller datasets typically available in pneumonia clinical research.

One of the most useful developments in the area of machine learning is causal forests and a limited number of other methods (Bayesian trees, Gaussian processes for machine learning, causal multivariate adaptive regression). These methods allow for the direct calculation of treatment effects, the effect of the predictor variable on the outcome. This can be particularly useful in the case of a negative study, when the results are inconclusive or suggest that the intervention is not effective. In these negative studies, there may be subpopulations or clusters of individuals who will have a different and sometimes clinically meaningful treatment effect. Some subpopulations may even result in opposite associations between predictor and outcome. This differential treatment effect among subsamples is termed 'heterogeneous treatment effects', and can now be identified using machine learning methods. Although one can use traditional methods to identify subpopulations through stratified regression modeling or inclusion of interaction terms (to detect effect modification), the investigator will run into issues with multiple testing bias, a common pitfall in frequentist statistics. In those hypothesis testing-driven methods, each

***Correspondence To:** Timothy L Wiemken, PhD
Work Address: Saint Louis University Center for Health Outcomes Research
(SLUCOR) 3545 Lafayette Ave #411 St. Louis, MO 63104
Work Email: timothy.wiemken@health.slu.edu

Table 1 Summary of methods

Method	Description
Frequentist Statistics	
Hypothesis Testing	Input data either rejects a null hypothesis or not, usually using a 5% level of significance
Machine Learning	
Decision Trees	A tree structure of decisions regarding input data is created and paths can be followed to make a prediction
Support Vector Machines	Input data is separated into 2 classes to make predictions and generalizations about each class
Artificial Neural Networks	A series of connections modeling the human brain are given input data and trained iteratively to make predictions
Deep Learning	Many more layers of connections are used to train the neural network to increase the accuracy of its predictions

Figure 1 Sample Decision Tree

additional test run on the data increases the amount of statistical error present. Therefore, if an investigator wishes to evaluate ten different variables as potentially providing different effects among the study sample, the level of statistical error increase substantially. At the most conservative, one can use the Bonferroni correction to account for this. In this approach, the cutoff for statistical significance (again, typically 0.05) is divided by the number of tests run, allowing for identification of a new cutoff. Here, 0.05/10 = 0.005. Therefore, no variable in the models could be considered significant until the P-value is below that new threshold. Given the typical sample sizes in pneumonia clinical research, this is unlikely to occur, even if there is a true and large treatment effect in the data. Our novel machine learning methods do not suffer from this issue, as they are not focused around hypothesis testing, rather algorithmic approaches to defining treatment effects. Because of this, one can evaluate as many variables as they wish for defining heterogeneous treatment effects without increasing error.

Machine learning algorithms

The term machine learning is an umbrella term comprising multiple computational methods. These methods are often combined with traditional frequentist statistical approaches such as bivariable tests, basic general linear models including logistic regression, leading some individuals to prefer the phrase "Statistical Learning". For the purposes of this paper, we will use machine learning.

One can think of machine learning as a set of tools used to algorithmically compute a prediction for a clinical outcome. Often the outcome is binary, such as the presence versus absence of a disease or mortality versus survival; however, an outcome also can be multiclass such as a risk score like the Pneumonia Severity Index [2].

Logistic regression is widely used for analysis in clinical research. However, it suffers from several problems including difficulty with interpreting log-odd or odds ratios, comparing log-odd ratios calculated with different independent variables, and comparing log-odds and odds ratios across subgroups in the sample, even if they use the same independent variables [3]. Several machine learning algorithms are available that can also be used that overcome some of the problems of logistic regression. These include, but are not limited to, Decision Trees, Support Vector Machines, Artificial Neural Networks, and Deep Learning (**Table 1**).
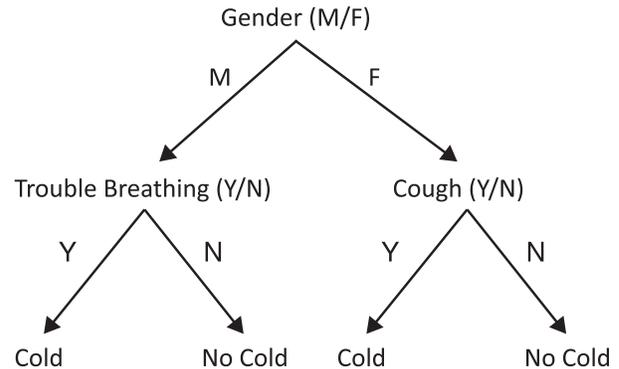
Decision Tree

Decision tree algorithms determine how to split a dataset into a tree structure that can be followed from the top of the tree to the bottom to make a prediction. Decision trees are already widely used in epidemiological and clinical research. For example, consider a small dataset with variables gender (M/F), trouble breathing (Y/N), and cough (Y/N); the outcome is whether or not a person has the common cold. A decision tree algorithm uses the dataset to "learn" the rules and build the tree for predict the outcome (**Figure 1**).

Several algorithms exist for creating decision trees including ID3 [4], C4.5 and CART [5].These approaches have been further enhanced to reduce error in the computation with ensemble decision trees in which multiple trees are generated from the data and their results averaged. Examples include random forests, extreme gradient boosted machines, and causal forests. In fact, CART and random forests can be used for propensity score generation to overcome some of the limitations of logistic regression [6].

Support Vector Machine

A Support Vector Machine is a machine learning technique that seeks optimal separation between two classes of data. SVMs calculate a margin, which is the largest region that separates two classes (**Figure 2**). The margin is also called the separating hyperplane. Imagine using a piece of paper placed on its edge on the line in **Figure 2**. The paper would cleanly separate these classes with the margin being the largest distance possible between the classes.
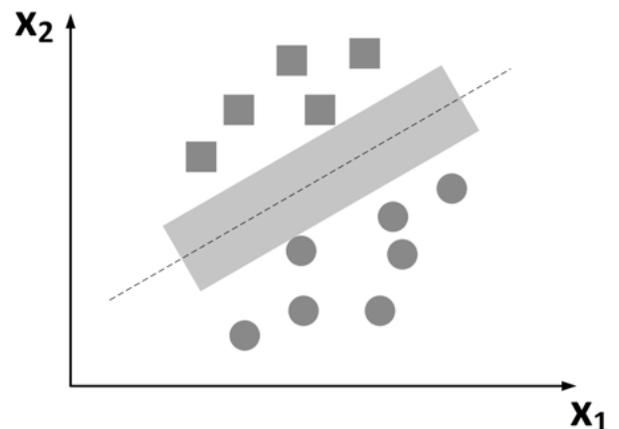

Figure 2 Support Vector Machine - margin

SVM works particularly well for problems that are linear, but can also be used on non-linear problems if they can be transformed into a representation that supports a separating hyperplane. For example, in **Figure 3-a**, no line can be drawn through the data to separate two classes. However, if the data were transformed into a "bowl" shape, a line can be drawn to separate them (**Figure 3-b**). The algorithm then can determine which class a data point belongs (**Figure 3-c**). The mathematical transforms to convert non-linear data into linear are called kernels. Several kernels exists for various linear and non-linear problems including, linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.
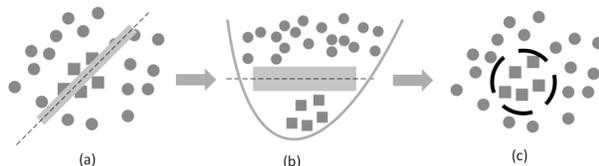

Figure 3 Support Vector Machine - transformation

One significant limitation of SVMs is that they are inherently binary classifiers. To accomplish multi-classification, they are typically run one against everything else and results of each classification compared. For example, if determining risk classification with four levels, SVM would classify the first level against all others, the second level against all others and so on. Like logistic regression, SVM also can be used to generate propensity scores. However, they are not as effective as boosting and decision trees [6].

Artificial Neural Networks
Artificial Neural Networks (ANN) model the human brain. The input data are represented as a set of nodes that connect to one or more hidden layers of nodes called neurons. The hidden layer then connect to an output layer that provides the result (see **Figure 4**).

The connections between each neuron (the lines in **Figure 4**) have weights that control whether or not a neuron "fires". Using a mathematical technique called gradient descent, the a neural network algorithm "learns" the optimal set of weights to fire the
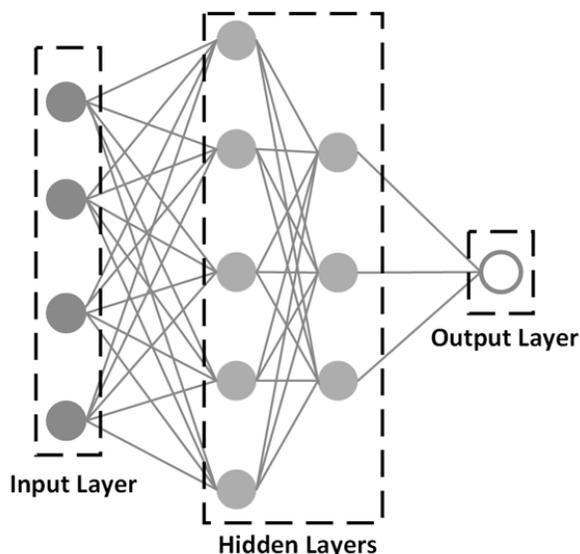

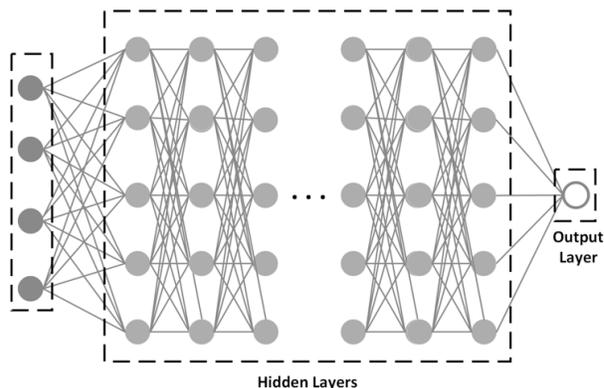**Figure 4** Artificial Neural Network

Hidden Layers
Figure 5 Deep Learning Artificial Neural Network

appropriate neurons that will discriminate between different classes by produce a 1 or 0 at the output layer. One advantage of ANNs over some of the other machine learning algorithms is that they train on the same dataset multiple times. Therefore, there is less of an effect on the training if the dataset is small. Furthermore, ANNs can model linear and non-linear problems. The are several other models of ANNs including recurrent neural networks in which nodes later in the model (toward the right side in **Figure 4**) can be connected to previous layers.

Deep Learning
An extension to the ANN is the Deep Artificial Neural Network often just called Deep Learning. Traditional ANNs only use a few hidden layer, usually ten or less. To model the human brain more effectively, the number of hidden layers can be expanded to several hundred layers. While this still doesn't approach the size of the networks in the human brain, deep learning has proven to be effective for many categorization tasks and pattern matching (**Figure 5**) [7].

The primary disadvantages of ANNs and Deep Learning is interpretability and overfitting. With linear regression and logistic regression, the coefficients can be interpreted, although as previously discussed, it is more complicated for logistic regression. With any form of neural network, the feature of the network that is optimized are the weights between the neurons. There is no straightforward way to explain how the weights affect the prediction. In contrast, I decision tree, even if large, has paths that can be followed to show how the outcome is derived. Overfitting, in which the ANN learns the training data extremely well, but does not generalize to other data is another significant problem. This can be addressed with regularization which is a common mathematical technique that adds data to the model so it better generalizes to other data.

Conclusions

When analyzing an association in clinical research, our traditional statistical methods, or frequentist statistics, are too rigid. The results are either significant or nor significant and the study is either positive or negative. One important advantage of machine learning will be our capability to better evaluate effects of interventions in study subpopulations. We think that the adoption of machine learning in clinical research will open a new and more productive way to analyze study results. Our capacity for interpretation of data and the generation of study conclusions will be enhanced by machine learning.

References

1. Furmanek S, English CL, Chandler T, Wiemken TL. Most Common Statistical Methodologies in Recent Clinical Studies of Community-Acquired Pneumonia. Journal of Respiratory Infections. 2017;1(4). https://doi.org/10.18297/jri/vol1/iss4/9.

2. Fine MJ, Hanusa BH, Lave JR, Singer DE, Stone RA, Weissfeld LA, et al. Comparison of a disease-specific and a generic severity of illness measure for patients with community-acquired pneumonia. J Gen Intern Med. 1995 Jul;10(7):359–68.

3. Mood C. Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur Sociol Rev. 2010;26(1):67–82.

4. Quinlan JR. Induction of decision trees. Mach Learn. 1986 Mar;1(1):81–106. https://doi.org/10.1007/BF00116251.

5. Breiman L. Classification and Regression Trees. Belmont (Calif.): Wadsworth International Group; 1984.

6. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. J Clin Epidemiol. 2010;63(8):826.

7. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015 Jan;61:85–117.