Peer Reviewed

# Machine Learning Workflow – Part 1

*John B. C. Tan, PhD, Fu-Sheng Chou, MD, PhD*

Starting this month, we would like to go through the entire workflow for developing a machine learning model and highlight some of the available techniques for each step. The goal is to introduce terminology and provide a general idea about the workflow rather than in-depth technical details. We hope you find this article helpful if you are thinking about developing one or two prediction tools for your practice. We are also open to collaboration to help you get your first predictive modeling project started.

***"This month, we would like to take a deeper dive into discussing two major types of classical machine learning and introducing several "non-linear" algorithms to the readers."***

## Overview

Figure 1 depicts the general process for machine learning model development. Data collection and processing are key to successful modeling and will be discussed in this article. In predictive modeling, data is typically split into training and testing datasets. While the training dataset is used for model development and fine-tuning, the testing dataset should not be touched. The testing dataset is used to validate the model externally. Ideally, in the context of healthcare predictive modeling, the training data set should be collected retrospectively, and the test data set should be collected prospectively. However, this assumes that the prospective test cohort will follow the same standardized and stringent clinical management standards as the retrospective training dataset. More frequently, the retrospective dataset is split into the training and the testing datasets. There will be more discussion about ways to construct the training and testing datasets in a later article.

## Data preparation

*Define the question and determine the prediction goal*

In the practice of neonatology, we deal with uncertainty about the future constantly. Now that the mortality rate of neonates has decreased, the current focus of neonatal research has shifted from "how can we get these babies to survive?" to "how can we give these babies a good quality of life?" While this goal may seem too ambitious at first glance, we can utilize the latest machine learning algorithms and predictive modeling practices to contribute significantly to the answer to this question. In general, the goal of predictive modeling is to use the hidden patterns in current or past data to predict the outcome. "Hidden," in this instance, refers

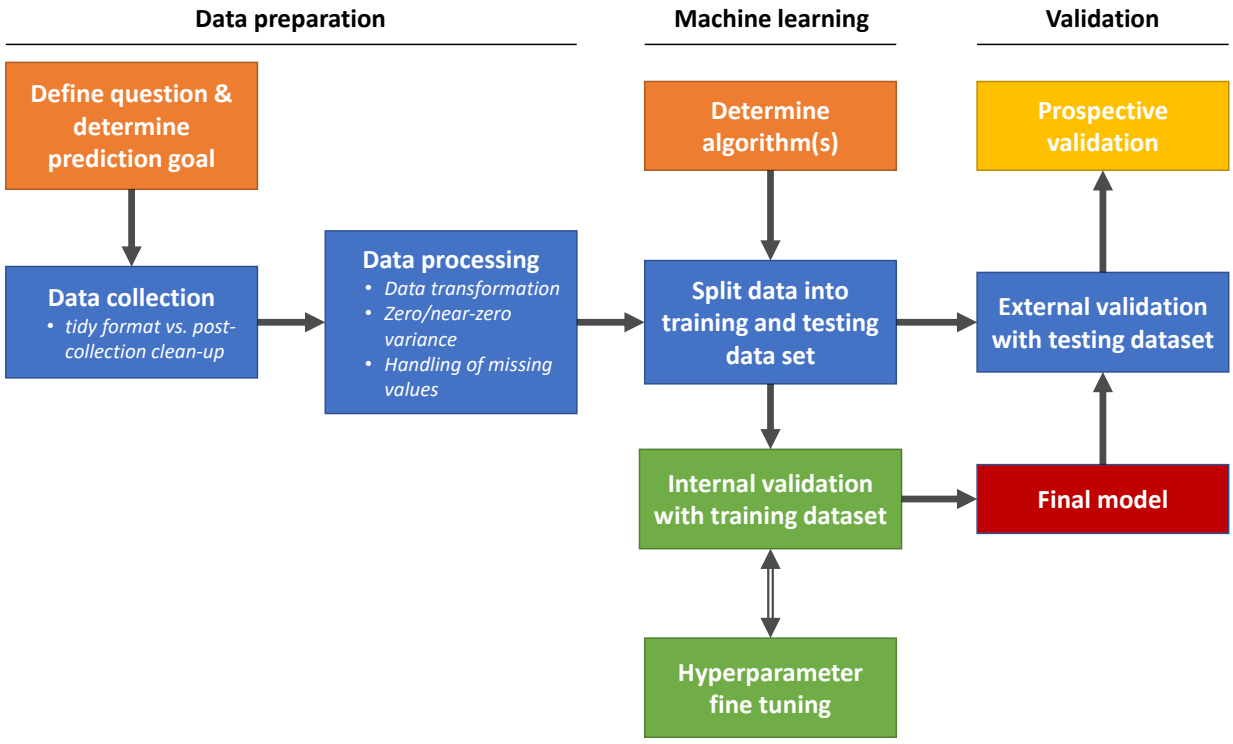



*Figure 1: The general process for machine learning model development.*

to the fact that there are a lot of quantifiable patterns in biology that are unrecognizable to the human eye; therefore, we must rely on machines to recognize these patterns for us. While traditional hypothesis generation and testing may not always be necessary for a predictive modeling project, as such projects are exploratory by nature, it is imperative to identify the clinical question and the prediction goal. Key aspects to consider are (1) cohort size and (2) clinically relevant variables.

Cohort size is critical to successful modeling. Machine learning algorithms require a large amount of data to sort out the patterns. The orders of magnitude for the amount of data required depend on the type of machine learning algorithm. Classic machine learning relies on data points in the range of at least hundreds, while deep learning requires data points in the range of thousands. Furthermore, there may not be enough data within a single center to successfully create a prediction model depending on the prediction goal. Therefore, it is essential to determine whether data from a single center will suffice or if a multi-center collaboration is required.

***"Classic machine learning relies on data points in the range of at least hundreds, while deep learning requires data points in the range of thousands. Furthermore, there may not be enough data within a single center to successfully create a prediction model depending on the prediction goal. "***

Using clinically relevant variables is also key to a successful predictive modeling project. We are trained to use an existing staging/grading system to "quantify" disease severity in clinical practice. The classification system usually has a prognostic association but is not "predictive." For example, a recent report using data from Vermont Oxford Network showed that Grade 3 bronchopulmonary dysplasia is associated with a significantly higher likelihood of having a tracheostomy (18.3% vs. 1.0% in all infants) and longer hospital stay (median discharge at 49 weeks postmenstrual age vs. 39 weeks in all infants) (1). However, there is no guarantee that using this existing classification system is the best way to capture the hidden patterns within our data. Depending on the project's goal, the variables chosen for your predictive model should be intentional and discussed thoroughly. Notably, in machine learning jargon, "variables" are known as "predictors" or "features."

*Data collection*

Once the question has been defined, and a prediction goal is set, the next step is to collect data. We highly recommend collecting only raw data instead of processed or interpreted data. For example, collect data on PDA size in millimeters, not as small/moderate/large. Raw, untransformed, and humanly uninterpreted data with higher resolution will maximize the predictability. Plus, we can also recode the data elements into lower resolution as needed during data processing. It is also essential to use the tidy format in data collection: (a) each variable must have its own column; (b) each observation must have its own row; (c) each value must have its own cell (2). These rules seem straightforward, but it takes considerable effort when designing the data collection sheet to abide by them. One approach, which we have taken successfully, is to work with database experts in the medical informatics department to directly extract data from the database server that stores all the data for the electronic health records (EHR). While there is a slight learning curve when it comes to communicating with the database architect and learning about database systems, it does help tremendously to get the mindset of what tidy data means right after being familiar with the database system because all structured EHR data are stored in a tidy format. Notably, it is still tricky to extract free-text data from the EHR. Sometimes it is much easier to go into the charts and do a manual chart review. Ideally, data are collected in a tidy format. But if not, the subsequent data clean-up steps should be taken to restructure the data into a tidy format for more straightforward analysis.

Which specific data to collect and what format the data should be in is a very complex topic and can severely alter your predictive model outcomes. While it may not be desirable and is impossible to collect data on all variables, we do agree that we must bear an adventurous mind and be bold about what data to collect. To start, we may search the literature for risk and protective factors that have been reported to be associated with the outcome we would like to predict. These well-established risk factors can serve as "positive controls." For example, gestational age would be a good "positive control" when predicting respiratory outcome due to its strong correlation. Clinical observation is also important. We would suggest including those variables that may have a correlation with the outcome based on anecdotal experience. One may also include factors that may seem irrelevant to serve as "negative controls," although true independence between a feature and the desired outcome measure may not be detected by human perception. Collinearity is not a huge concern because there are techniques (regularization or non-linear algorithms as examples) to address it (3,4). For decision tree-based models, highly correlative variables should have similar importance scores (4). In such cases, the physiological relationship between the two variables in question can be further confirmed. Nonetheless, during the initial planning stages of the study, it is imperative to discuss which variables are to be collected and in which format. Deciding on these aspects of data collection midway through the study rather than the start of the study can lead to bias and overfitting.

*Data processing*

Data processing may include transforming data, eliminating zero- or near-zero-variance features, and dealing with missing values. Transforming data is particularly useful when using specific algorithms such as linear regression or Gaussian naïve Bayes, which require continuous variables to have a normal distribution. Exponential and power transformations (Box-Cox or Yeo-Johnson) are common techniques. Additionally, techniques such as centering around the mean and feature scaling may be performed to normalize or standardize the dataset. For example, if you have a variable with a range of 0 to 10 and another variable with a range from 0 to 1000, centering and scaling the data will strengthen your prediction model because it increases the comparability of the variables.

It is important to assess zero- or near-zero-variance in the features and eliminate these variables before training a model. Simply, when a variable has a high percentage of zero data, it is very likely that only observations with zero values will be selected when randomly selecting observations to construct the training dataset.

In this case, the variable becomes useless because all observations have the same value, which is zero. These predictors should be removed during data processing steps.

***"The major concern regarding missing values is whether the final model developed is skewed by the missing values and/or how the missing values are handled. Handling missing values by itself is a huge area of research and should be discussed with the machine learning engineer or the data scientist early on during project development."***

*Missing values*

Missing values are common in healthcare data and can be a headache in machine learning. While some algorithms, such as regularized linear regression and generalized additive modeling, can tolerate missing values, most algorithms require complete data. There are three major types of missing values: (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) (5). The mechanism of missingness in MCAR is independent of the collected or the missing data. In MAR, the mechanism of missingness is dependent on the collected data, whereas, in MNAR, the mechanism of missingness is dependent on the missing data. It is not easy to distinguish among different types of MCAR, MAR, and MNAR. The major concern regarding missing values is whether the final model developed is skewed by the missing values and/or how the missing values are handled. Handling missing values by itself is a huge area of research and should be discussed with the machine learning engineer or the data scientist early on during project development.

One way to deal with missing values is to eliminate the rows that contain missing values, but by doing this, the total observation number will decrease and may also cause bias in the final model. Generally speaking, if the missing data only comprise no more than 5% of the total observations, those incomplete observations can be ignored, an approach called complete case analysis (only analyzing complete data). On the other hand, if the percentage of the observations with missing values is significant, effort should be spent on understanding the missingness mechanism and whether another round of chart review or selecting additional surrogate features is feasible.

Alternatively, imputation of missing values may be performed. There are three "easy" methods to impute missing values: (a) add a constant value (using the last observed value or the worst value observed for the subject, for example); (b) use the corresponding value for the predictor from a random observation; (c) use the mean, median, or mode value from the variable. These methods can collectively be considered as single imputations. Single imputation has its own caveat, including underestimation of the variability and potentially creating bias. In contrast, multiple imputations, which are based on developing additional predictive models to predict the missing values, can be performed to impute the missing values. Multiple imputations are out of the scope of the article, so they will not be discussed here. Regardless of which method to choose, the same method must be used during the model validation process. Different methods can be tried to process the data for training. The best method may then be determined based on model performance, which we will discuss in Part 2 next month.

*Figure 3. Visual depiction of a decision tree.*

***"Regardless of which method to choose, the same method must be used during the model validation process. Different methods can be tried to process the data for training."***

**Summary**

Data preparation takes up most of the time during model development. Most data scientists would agree that they spend at least 80% of their time on data processing, emphasizing the importance of crafting well-processed data before training should occur. We hope you enjoy reading this article. We will discuss model training and validation next month.

***References:***


1. *Jensen EA, Edwards EM, Greenberg LT, Soll RF, Ehret DEY, Horbar JD. Severity of bronchopulmonary dysplasia among very preterm infants in the United States. Pediatrics. 2021 Jun 2;148(1):e2020030007.*
2. *Wickham H, Grolemund G. R for Data Science [Internet]. [cited 2021 Aug 13]. Available from: https://r4ds.had.co.nz/*
3. *Chou F-S. How are machine learning algorithms different from statistical methods? Neonatology Today. 2021 Jun 20;16(6):29–31.*
4. *Patel M, Tan JB, Chou F-S. Non-linear algorithms in supervised classical machine learning. Neonatology Today. 2021 Jul 20;16(7):40–3.*
5. *Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol [Internet]. 2017 Dec;17(1). Available from: http://dx.doi.org/10.1186/s12874-017-0442-1*



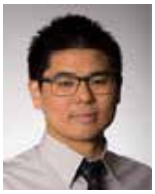
*Author contribution: Drs. Tan and Chou contributed equally to this manuscript.*

*Disclosure: The authors identify no conflict of interest*


**NT**

*John B. C. Tan, PhD*
*Data Scientist*
*Assistant Professor of Pediatrics*
*Division of Neonatology, Department of Pediatrics*
*Loma Linda University Children's Hospital*
*Email: JBTan@llu.edu*

*Corresponding Author*



*Fu-Sheng Chou, MD, PhD -*
*Senior Associate Editor,*
*Director, Digital Enterprise*
*Neonatology Today*
*Assistant Professor of Pediatrics*
*Division of Neonatology, Department of Pediatrics*
*Loma Linda University Children's Hospital*
*Email: FChou@llu.edu*