Machine Learning Workflow – Part 2

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

Last month, we discussed data preparation and processing. We introduced some terminology in data processing that you may encounter when reviewing or reading articles. Specifically, we discussed eliminating variables with zero or near-zero variance; we also briefly discussed ways to handle missing values. The most important and at the same time awkward terminology to keep in mind is, in the world of machine learning, variables are called "features"!

"This month, we will discuss more in-depth steps to take after initial data processing and the actual learning steps. Before we begin, we would like to point out that some minor changes were made to the machine learning flow chart (Figure 1) to align with the industry-level standard for machine learning in healthcare."

This month, we will discuss more in-depth steps to take after initial data processing and the actual learning steps. Before we begin, we would like to point out that some minor changes were made to the machine learning flow chart (Figure 1) to align with the industry-level standard for machine learning in healthcare. In this revised flow chart, internal validation refers to using institutional datasets (training and testing) for data validation. External validation indicates taking data from a different environment (extramural, multi-center, etc.) to validate the model further. On the other hand, prospective validation means further validation of the model using prospectively collected new data. We will discuss this in more detail later.

Considerations in splitting data for model development

Now you have collected all the raw data and gone through the painful processing steps to put all the data into a nice tidy format. You may have data from one NICU for one year or data across multiple years from multiple NICUs that are all managed under the same protocol by the same group of neonatologists. Now the question is, do we use all of the data for model development? A short answer is NO. A typical approach is to split the data 80:20 and take the larger portion for model development. But how do we decide how to split the data?

Random split

This is the most straightforward way of doing the split. All the data

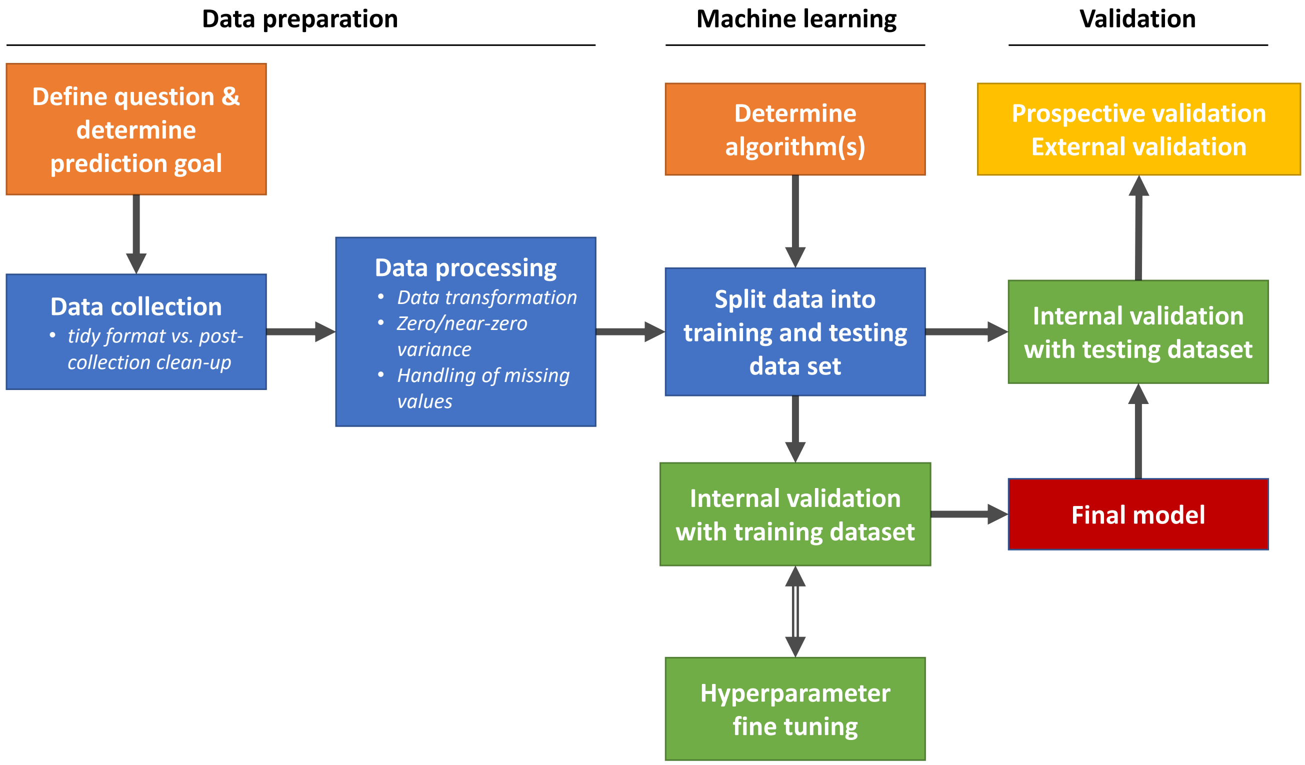

Data preparation
Machine learning
Validation
Define question & determine prediction goal
Data collection
• tidy format vs. post-collection clean-up
Data processing
• Data transformation
• Zero/near-zero variance
• Handling of missing values
Determine algorithm(s)
Split data into training and testing data set
Internal validation with training dataset
Hyperparameter fine tuning
Final model
Internal validation with testing dataset
Prospective validation
External validation

Figure 1.

are lined up and randomly split into two groups based on the outcome feature we would like to predict. Classic machine learning algorithms cannot consider repeated measures, like what mixed modeling can provide. Therefore, features such as year, location, or patient ID, intended to address repeated measurement in the data may become an issue for future prediction, as those features need to be supplied in future data, which may sometimes become awkward. For example, if one includes data from the year 2010 to 2015 for modeling, and a categorical feature "Year" was included in the training dataset. Now, you have a set of data from 2021 that you would like to supply to predict the outcome. You would then supply 2021 for the feature Year, but it would not make sense to the model. This is because 2021 was never observed during model development as a possible value for feature Year. It would also not make sense to treat Year as a continuous data type.

> ***"When multiple years of data are involved, it usually means that the study is retrospective in nature, which is typical for machine learning projects, in healthcare, or anywhere else, as one of the intrinsic nature of prediction is 'to learn from the past.'"***

Spatial split

When data are obtained from more than one NICU, one can either randomly take data from all NICUs for model development and use the remaining data for validation or take data from some designated NICUs for model development, then validate the model using data from the remaining NICUs. The first approach is like what we just discussed in the *Random split* section. It is intuitive and makes perfect sense. On the other hand, the second approach may be appealing for several reasons:

1. Logistic considerations:

   If each NICU requires an independent ethics review (IRB) process, it will take a long time to get all applications reviewed and approved from all institutes. It would also lengthen the time needed to gather raw data from all NICUs for processing and randomization. This may not be feasible from a project development and funding standpoint. One may consider taking data from 1 or 2 NICUs to establish the data organization and processing pipeline and use these data for model development. Findings obtained from this process may be turned into a grant proposal to secure more resources for model fine-tuning and external validation.

2. **Clinical practice comparison:**

   In clinical protocol (the flowcharts with the boxes and the arrows that we, at some point, have all been involved in developing) implementation, we tend to test the protocol in one unit for a defined period of time. This allows us to fine-tune the protocol to address any obstacles that may be encountered during implementation. After successful implementation in a confined environment, we would then expand it to other units. There may be additional obstacles that are NICU-specific that need to be further fine-tuned. We can take a similar approach to machine learning model development. After all, one common goal of developing prediction models is to allow the machine (the algorithms) to tease out the hidden patterns in the data to inform the relationship between the features and the outcome. We can take data from 1 or 2 NICUs to develop a prediction model. We investigated the model well to understand the essential features that the machine had learned using a chosen algorithm. We can then try to validate the model using data from a separate environment and assess whether the model still holds. We may also develop two models using data from two different sources and compare the list of important features between the two models. Either way, such an approach provides an opportunity for us to understand the difference in practice between locations.

3. Identification of essential predictive features:

   Careful selection of features for model development is key to successful model development. Features that are specific to one NICU may not be applicable to another NICU. For example, if one NICU only uses a high-frequency jet ventilator (HFJV) for rescue, and the other NICU only uses a high-frequency oscillator (HFO) for rescue, including a feature of whether HFJV is used to train a model using data from the first NICU is not going to result in a generalizable model. If we perform careful feature selection and develop a generalizable model for another NICU, we know these features will be critical to all practices.

Temporal split

When multiple years of data are involved, it usually means that the study is retrospective in nature, which is typical for machine learning projects, in healthcare, or anywhere else, as one of the intrinsic nature of prediction is "to learn from the past." We typically include multi-year data for different reasons: we may not have enough data from just one year. Also, we may include a specific period because of protocol change or because a new initiative was started to improve care for a specific patient population. Comparisons between epochs in traditional statistical analysis are sometimes made for retrospective data because there was no good control group for the clinical question, or the clinical question was geared towards understanding how clinical outcomes evolve over time. There are a few considerations when it comes to splitting the data based on time:

1. Evolvement in clinical practice:
   It is essential to know whether the clinical practice has evolved during the period data were collected. For example, our NICU used to use HFO as a mode of ventilation for extremely low gestational age newborns (ELGANs), but the practice has moved away from it, and now HFJV is almost exclusively used for this population if a high-frequency mode of ventilation is needed. Having a feature that indicates the use of HFO would yield a near-zero answer in the more recent cohort and lead to significant errors in the model. On the other hand, if two features were included, one for HFO use, and the other for HFJV use, the model may then provide an opportunity to assess the difference between HFO and HFJV in predicting the outcome of interest. However, the conclusion may be misleading if the temporal effect is substantial since the differences may be due to time rather than the ventilation method. It is vital to take into account changes over time to reduce confounding as much as possible.

Table 1. Supervised learning algorithms and their strength and weakness

	Use	Principle	Strength	Weakness
Support vector machine	Classification, regression	A "maximum distance" approach to creating a hyperplane which provides the largest separation between outcome classes in a high-dimensional space	1. High number of features 2. Less likely to overfit 3. Small training dataset 4. Fast	1. High demand for computing power and computer memory 2. Difficult to interpret the relationship between input features and output.
Naïve Bayes Classifiers	Classification	Based on Bayes' probability theorem. All features are presumed to be independent of others.	1. Fast 2. Can tolerate a high number of features 3. Does not require a massive number of training data	1. Less accurate, especially with a low number of features (low-dimensional data) 2. The assumption of total independence of all features
k-nearest neighbor	Classification, regression	Use a decision boundary based on the hyperparameter *k* to determine the class of unknown data points, e.g., if *k*=5 and 3 out of 5 neighbors are *positive*, then the unknown is assigned as *positive*.	1. Simple to interpret 2. Good for data that has no prior knowledge about its distribution	1. Not suitable for data with a high number of features (high dimensional data) 2. Cannot tolerate missing values 3. Likely to overfit with high-dimensional data
Random forest	Classification, regression	A decision-tree-based approach with nodes and branches by "planting" a predetermined number of trees to avoid overfitting and high sensitivity to subtle changes in the training data. Handles missing values differently than single decision trees and does not prune the trees.	1. Easily explainable and excellent graphic representation 2. Tolerate missing values by conducting imputation or taking a proximity-based measure 3. High performance	1. High demand for computing power 2. Slow
Panelized regression	Classification, regression	In addition to linear regression, penalizing features to shrink large coefficients (Ridge) or drop less important features (LASSO) Elastic net regression is a mixture of ridge or LASSO regression, and in theory, is superior to either regression alone	1. Address overfitting issue with linear regression to provide better generalizability 2. Fast	1. Does not tolerate missing data 2. Not useful in data with a non-linear relationship

2. The temporal feature:
   To reiterate, it is not advisable to include the temporal feature in the data for modeling, as including a temporal feature that is not recurring (e.g., the year is not recurring, the season is recurring) will make the prediction or future events impossible. It is also important to be cautious when including features that are highly correlated with time, as it may give a false impression about the importance of the feature. Readers interested in temporal effects may wish to read more on the concept of time-series forecasting.

Choosing machine learning algorithms

In the past articles, we introduced supervised vs. unsupervised algorithms and linear vs. non-linear algorithms. As most of what we try to accomplish involve predicting the outcome, we typically deal with supervised learning. When choosing algorithms, it is essential to consider sample size, missing values and how they are handled, and the feature size. For example, a support vector machine can tolerate a smaller sample size and high-dimensional (a lot of feature numbers) data. On the other hand, a random forest cannot tolerate missing values and cannot extrapolate. The advantages and disadvantages of some of the common algorithms are listed in Table 1. While this article does not intend to provide an exhaustive list of all the strengths and weaknesses of all available models, we advise the readers to conduct online searches on the characteristics of algorithms they encounter to understand better the suitability of the algorithm for the clinical questions. It is also essential to discuss which algorithm(s) to choose to build the prediction model with the machine learning engineer. Key aspects to ask include:

1. Handling of dimensionality (high vs. low number of features)
2. Classification vs. regression
3. Training data size
4. Missing value tolerance
5. Real-time modeling (training time) and computational costs
6. Tendency for overfitting
7. Accuracy

> ***"Finally, the metrics for performance assessment are also something to discuss with the machine learning engineer and will be discussed in a later article."***

Finally, the metrics for performance assessment are also something to discuss with the machine learning engineer and will be discussed in a later article. Sometimes it may be desirable to trial different algorithms empirically, and that is okay. Machine learning projects are, after all, not hypothesis-driven studies. A predetermined analytic approach, the required sample size for adequate statistical power, and power analysis are usually not considered. The ultimate goal is to create the best model that gives the best prediction performance.

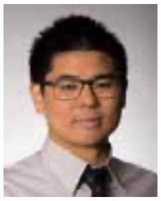
Author contribution: Drs. Tan and Chou contributed equally to this manuscript.

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