Paper ML07

Data Review of RECIST Criteria using ML/AI
A Proof of Concept Example

Priya Gopal, TESARO, a GSK Company, Waltham, MA USA
Christine Teng, Merck & Co., Inc., Rahway, NJ USA
Sam Tomioka, Sunovion, Fort Lee, NJ USA

ABSTRACT

Machine learning (ML)/Artificial Intelligence (AI) provides the ability to learn from data patterns and make decisions without the need for explicit programming. In oncology clinical studies, data management/medical review team verifies the tumor measurements and corresponding assigned RECIST criteria using manual approach to determine quality of the response data entered by the sites. Depending on the size of study population and complexity, the review process may involve additional resources and time to check for each data point in the study.

This paper focuses on a proof-of-concept (POC) of how machine learning (ML) algorithms/Artificial Intelligence (AI) could help increase efficiencies with review and cleaning process of tumor data. Realistic simulated data containing tumor measurements and corresponding RECIST response values are synthesized and used to train the various algorithms and neural networks in identifying the data patterns to infer RECIST criteria. Data was divided into training and test sets to determine the accuracy of predictions made by selected supervised ML algorithms. This POC is developed in using scikit-learn and TensorFlow framework.

Key Words: Machine learning (ML), Artificial Intelligence (AI), RECIST, Oncology, Python

INTRODUCTION

The Response Evaluation Criteria in Solid Tumors (RECIST) provide standardized rules for solid tumor response assessments. RECIST defines categories of response, which include Complete Response (CR), Partial Response (PR), stable disease (SD), and progressive (worsening) disease (PD), after the start of treatment.
There are three types of lesion in the overall tumor response assessment by RECIST. Target lesions are defined as those followed quantitatively. Non-target lesion as defined as those followed qualitatively as a whole. New lesions are evaluated in a qualitative and binary fashion in typical RECIST assessment, and the appearance of a new lesion is interpreted as progression. Site investigator assessments of the target, non-target, and new lesion status, along with an overall response assessment, are entered for each tumor assessment time-point. Variability is typically greater in site review than in central review, because site reviewers are a larger group with less consistent training in applying formal criteria than the reviewers at dedicated independent review facilities. However, both Site and Central review should follow the same criteria in terms of definitions and methods.

Response Criteria

Target lesions (TL) are measured at every visit, and their measurements are added to get the sum of diameters. Non-target lesions are evaluated as ‘present’, ‘absent’ or ‘unequivocal progression’. A lesion identified post-baseline is consider a new lesion, which indicates disease progression.

The focus of this POC exercise does not consider evaluation of Non-target lesion in the data pattern. The timepoint is not considered either as the part of the data pattern so will be excluded as well. Several supervised ML algorithms are used to train and test the data to predict the response.

Target Lesion (TL) Response Criteria

To derive target response, we need to have the following information:

- Identify Longest Diameter (LD) for non-nodal target lesions

- Identify Short Axis (SA) for target lymph nodes
- Add up Sum of Diameters (SOD) from LD and SA
- Calculate percentage change from nadir SOD
- Calculate percentage change from baseline SOD

Please note that the “nadir” is the smallest SOD prior to the current assessment.

Follow below criteria to derive the target lesion time-point response. If an assessment met multiple criteria, PD overwrites all response result and NE overwrites any result other than PD. With any new lesion at post-baseline, response is PD.

Response	Definition
Complete Response (CR)	All non-nodal TLs disappeared; all lymph nodes short axis <10 mm
Partial Response (PR)	SOD decreased ≥ 30% from baseline
Progressive Disease (PD)	SOD increased ≥ 20% from nadir and the 20% has absolute increase ≥ 5 mm
Stable Disease (SD)	Not PR nor PD
Not Evaluable (NE)	Cannot determine target lesion response

SITE vs. CENTRAL REVIEW

There are several advantages using assessments from central review, including reads performed by a small group of highly trained readers, a controlled read system, and complete blinding to treatment group and clinical condition. The RECIST derivations can apply to both Site and Central Review data to confirm the quality of the response.
As mentioned earlier, certain level of discordance between site and central results is expected, and the degree of discordance is not systematically described in the scientific literature. The discordance could be summarized at different levels, e.g., for each target lesion/non-target lesion/new lesion response, overall response at each subject visit, or best overall response at subject level. These summary reports provide more information than the discordance report at study level. The summary reports can detect both the systematic errors in site imaging review and potential data issues due to other reasons. Data management and clinical operation need to understand the causes of discordance, and the report of site vs. central discordances can be an important source of information.

Since central data is managed by an external vendor before loading into clinical database, data points and issues were addressed prior transferring to the sponsor. However, site data is collected from Case Report Forms (CRF) and often we found missing data points which may affect the derived result.

WHAT IS ML and AI

Artificial Intelligence (AI) is Intelligence demonstrated by machines, in contrast to intelligence displayed by humans. Machine Learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

There are three major recognized machine learning categories: supervised learning, unsupervised learning, and reinforcement learning.

1. **Supervised learning algorithms** are used when the data contains input and set of desired output (labeled data). Data classification algorithms and regression algorithms are part of supervised learning model.

2. **Unsupervised learning algorithms** are used when the machine is trained using unlabeled data. The data is grouped or clustered into different categories to find structure in the data.
3. **Reinforcement learning algorithms** is different when compared to supervised and unsupervised learning. It continuously learns from the environment in an iterative fashion.

Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input feature*s* so that we can predict the output values for new data based on learned relationships from the training data sets. For this POC, we are using supervised ML algorithms to predict the RECIST criteria.

RESPONSE CRITERIA PREDICTION FOR TUMOR

The input data were prepared with following columns in numeric format and missing values were imputed before process. The imputation methods are discussed later.

Data structure

- BSUM – Baseline Sum of Diameter
- SUMDIAM – Sum of Diameter (Sum of Longest Diameters and Sum of diameter of non-lymph-node tumor) at current visit
- PCBSD – Percent Change From Baseline in Sum of Diameter
- NADIR – smallest Sum of Diameter prior to current time point
- ACNSD – Absolute Change From Nadir in Sum of Diameter
- PCNSD – Percent Change From Nadir in Sum of Diameter
- NEWLSN – New Lesion (1-yes,0-no)
- TRGRESP – Target Lesion Response CR, PR, PD, SD, and NE (label)

The tumor prediction criteria were created using site and central data. There are missing data value in one or more variables in the above list for some subjects in the site dataset. Initially, the datasets were prepared and trained as follows: These scenarios were for exploring purpose to evaluate the ML algorithms.

	Training and Validation Data	Test Data	Test Data ID
1	Central	site	A
2	Site	central	B
3	central*85%+site*85%	central*15%	C
4	central*85%+site*85%	site*15%	D

Since the central data are used as the primary efficacy endpoint, later the training, validation and testing were performed as follows. The idea is that the model trained on the central can be used to identify potential inconsistent assessment of the response criteria, therefore, the assumption is that the accuracy of the model using the site is expected to be lower.

	Training and Validation Data	Test Data	Test Data ID
5	central*70%	central*30%	E
6	central*70%	site	F

Following algorithms were used to generate predictive models in this paper.

- K Nearest Neighbor
- Random Forest
- XGboost Gradient Boosting
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks)

Previously, Atbour et.al. reported at ASCO the 4-layer neural network architecture (ml-RECIST) can accurately estimate the outcomes using imaging text report. Here we generated 4 different architectures to evaluate the performance including ml-RECIST.

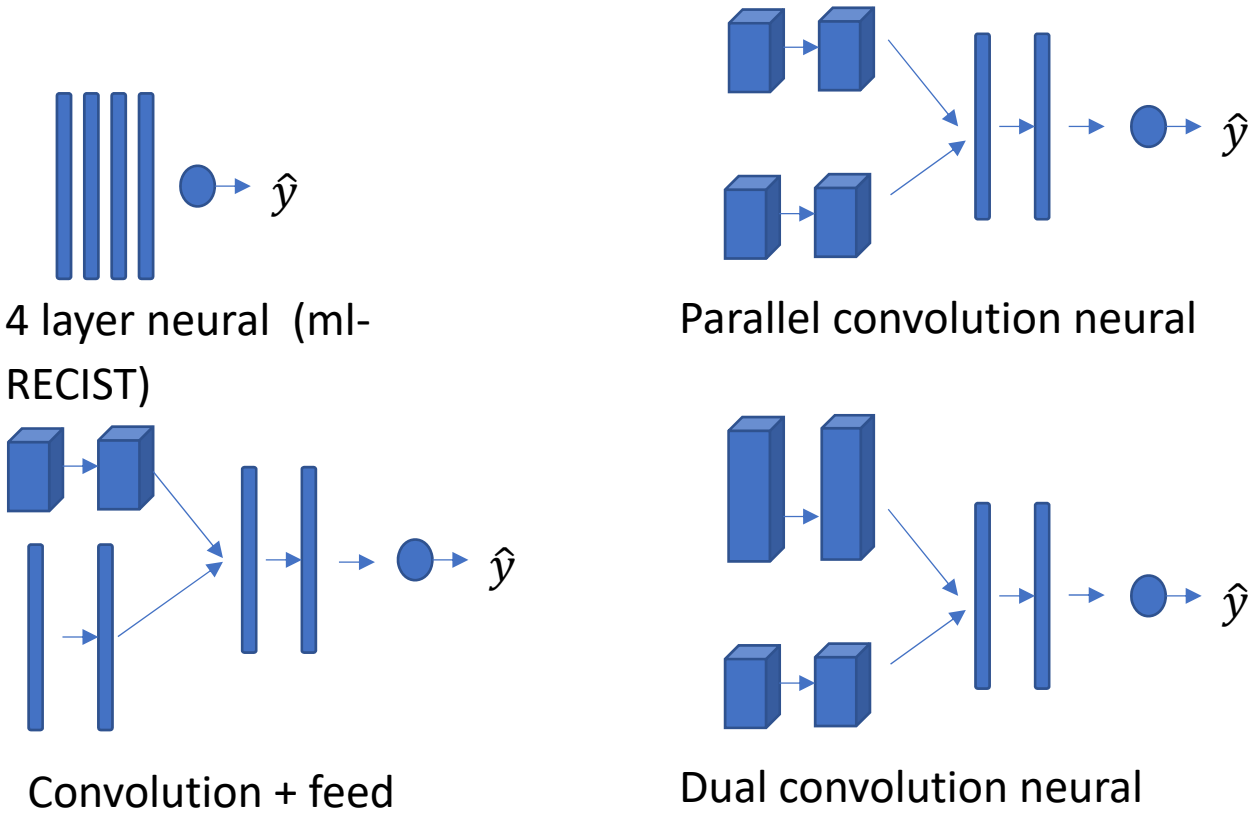

Figure 1: Illustration of neural network architecture trained and tested

Stratified split was performed to generate 20% of data for the testing, and 80% was used for the training and validation. Except for Neural Networks algorithms, stratified 3-fold cross validation was done during the training to verify the performance of the models. The model for each algorithm was selected based on the performance of the cross validation.

80%		20%
2/3	1/3	
Training	Validation	Test

For the neural network models, 80% of the data were used to split for training 80% and validation 20% as illustrated in the following figure.

80%		20%
80%	20%	
Training	Validation	Test

The entropy loss was used to optimize the neural network models. Given the training data X, the entropy loss function is defined as

$$H(p,q) = -\sum_{x \in X} p(x) \log q(x)$$

where q is the predicted distribution and the q

IMPUTATION FOR MISSING DATA

Site data is collected and entered by different sites. Sites often have missing data so we will need to impute some missing data in order to use ML algorithms. XGboost doesn't require missing data imputation, but we used the imputed data for the training.

Imputation of NADIR
If NADIR is missing, we assumed that no post baseline measurement was collected, therefore BSUM was used to impute NADIR.

Imputation of PCBSD and SUMDIAM
When TRGRESP is NE, PCBSD may not necessarily be missing, but when PCBSD is missing, the TRGRESP was always NE. Missing PCBSD depends on missing SUMDIAM. Missing SUMDIAM was assumed missing completely at random.

Following imputation steps were taken for PCBSD and SUMDIAM.

- If SUMDIAM is not missing, PCBSD was derived as the difference between SUMDIAM and BSUM over BSUM.
- SUMDIAM imputed value was dynamically chosen based on the minimum loss of the model. We used h to denote it as one of the training hyperparameter for the model optimization. This is the stand alone hyperparameter we added in all of the ML algorithms and neural network models we trained. We defined the range of *h* prior to train the models and choose the best *h* based on the performance metric. The objective of the model is not to predict the RECIST criteria when SUMDIAM is missing but classify it as NE.

Imputation of NEWLSN
Missing NEWLSN was set to 0.

Additional imputation methods
Additional imputation methods were explored to confirm the generalizability of the models based on above imputation methods.

- Median imputation for continuous variables, set to 0 for NEWLSN
- K-nearest neighbor
- Gradient boosting
- Stochastic gradient descent
- Naive Bayes
- Decision Tree

ML RESULT COMPARISONS

	Training and Validation Data	Test Data	Test Data ID
1	Central	site	A

2	Site	central	B
3	central*85%+site*85%	central*15%	C
4	central*85%+site*85%	site*15%	D
5	central*70%	central*30%	E
6	central*70%	site	F

Table 1 training, validation, and the test data descriptions

	Test Data ID	metric	rf	svc	lr	knn	Xgb
1	A	acc	90.1	84.5	88.1	84.6	90.4
2	B	acc	83.0	73.9	78.0	81.1	82.3
3	C	acc	82.6	73.9	81.2	82.6	84.1
4	D	acc	95.2	86.5	90.3	93.3	95.2
5	E	acc	84.4	72.1	81.5	85.9	85.9
6	F	acc	88.4	82.6	87.9	87.4	89.2

Table 2 Results from the various classifiers (macro accuracy)

	Neural network	Test Data ID	Acc
7	ml-RECIST	C	86.2
8	ml-RECIST	D	97.1
9	fnn+ convolution	C	87.0
10	fnn+ convolution	D	97.1
11	Parallel Convolution	C	**86.2**
12	Parallel Convolution	D	**98.0**
13	Dual Convolution	E	85.8
14	Dual Convolution	F	92.0

Table 3 The results from the neural network based models (macro accuracy)

In practice, the model will be used to check the site data, hence the results on test data A and D are close to the real implementation of this approach. The scenario #1, 4, 6, 8, 10, and 12 suggest that the model for validating the new site data, the training data should contain the data from the central and the site and since the explanation for the identification of issues may not be important here. While previously proposed ml-RECIST performed well on Test Data C and D the parallel convolution approach archived the best performance.

SUMMARY

Five algorithms and 4 different architecture of neural network were tested to predict RECIST responses including not evaluable. The neural network models consistently outperformed the machine learning algorithms.

Various imputation values were used for SUMDIAM to train the models and confirmed that the imputation values do affect the performance of the models. In the practice, it is recommended to include the missing data imputation as part of the training to learn the best imputation for the model performance as we implemented for this paper.

We also tested various scaling methods to potentially optimize the computation speed, but the no scaling and the robust scaling tend to produce better results for the ensembled models.

The recommended approach to construct the model is to use the mixture of central and site data, and construct the parallel neural network architecture or gradient boosting with XGBoost for the purpose of validating the site data.

REFERENCE

New Response Evaluation Criteria in Solid Tumors: Revised RECIST Guideline (Version 1.1), European Journal of Cancer, 45:228-247
https://ctep.cancer.gov/protocoldevelopment/docs/recist_guideline.pdf

Kathryn Cecilia Arbour, Luu Anh Tuan, Hira Rizvi, Adam Yala, Matthew David Hellmann, and Regina Barzilay. ml-RECIST: Machine learning to estimate RECIST in patients with NSCLC treated with PD-(L)1 blockade.

Journal of Clinical Oncology 2019 37:15_suppl, 9052-9052

ACKNOWLEDGEMENTS

We are part of PhUSE ML/AI Machine Learning Project Sub-team. We would like to thank our management's support and review of this paper; and thank PhUSE US Connect committee the opportunity to present.

TRADEMARKS

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sam Tomioka Sunovion Fort Lee, NJ USA Sam.Tomioka@sunovion.com
Christine Teng Merck & Co., Inc. Rahway, NJ 07065 christine_teng@merck.com
Priya Gopal TESARO, a GSK Company Waltham MA 02451 priya,x,gopal@gsk.com