Paper ET17

Case study of verification and validation methods for algorithmic predictive biomarkers

Kartik Shetty, Takeda, Boston, US
Priya Gopal, Takeda, Boston, US

ABSTRACT

Digital Health Technology (DHT) is an emerging field; known for its ability to make healthcare delivery to patients efficient. One application is the development of predictive biomarkers through the application of algorithms on data from wearable devices, enabling the prediction of health conditions; leveraging this potential can accelerate drug development. However, both data and algorithms need to meet regulatory requirements with regards to consistency, accuracy, traceability, and validity to be part of a clinical study. The focus of this paper is to investigate the verification and validation methods for algorithms that are utilized in the formulation of biomarkers. As a part of PHUSE "Emerging Trends & Technologies" working group's project "DHTs", we present a case study where heart rate variability based predictive biomarker is devised using an algorithmic approach. The objective is to develop a framework to guide the use of DHT in clinical trials for comparable scenarios.

INTRODUCTION

Biomarkers is a measurement captured at point in time on a cell or an organism whose presence is indicative of some phenomenon such as disease, infection, or environmental exposure. We are focusing on predictive biomarker which can predict a health condition ahead of time before it manifests. As digital health technology progresses, wearable devices have evolved to prioritize patient-centricity, offering enhanced ease of use and accessibility. These devices now efficiently gather extensive healthcare data for patients, enabling the utilization of software and technologies to generate predictive biomarkers. This generated predicted biomarker should meet clinical research requirements to be used in the clinical trial setting which in return can accelerate drug development and patient care.

We have developed a case study which uses the Welltory COVID-19 dataset to create a predictive biomarker using the Random Forest and Logistic Regression classification techniques. We will be using the case study to identify the steps we have taken to apply the algorithmic techniques and finally create the biomarker. We have attempted to provide recommendations and suggestions on how to verify and validate the biomarker to be compliant for clinical trial.

CASE STUDY

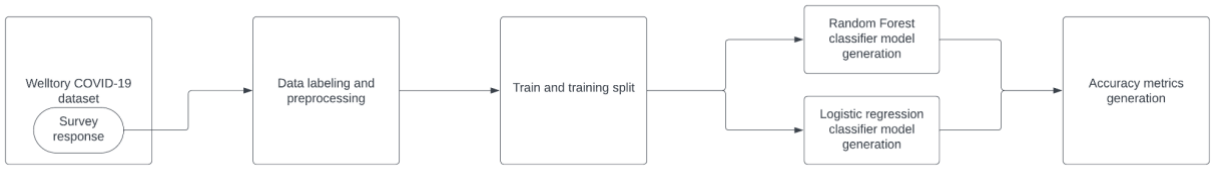
For the case study, the dataset was curated as part of the COVID-19 research which was carried out in 2020 by the Welltory team to detect patterns regarding the COVID-19 disease, progression and recovery. As part of the research, users with positive COVID-19 status had their symptoms, heart rate variability, and data tracked using the Welltory app.

The gathered information comprised measurements of heart rate variability, data from wearable devices such as Apple and Garmin watches, and clinically verified assessments of both physical and mental health, encompassing details about symptoms and test outcomes. The dataset contained data collected from 68 patients and had a size of 24 megabytes.

We wanted to create a predictive biomarker to anticipate severe COVID symptoms in users who tested positive, using data points obtained from the patient's wearable devices. By leveraging survey responses that assessed the user's symptoms, we labeled the wearable data points to differentiate between cases with severe COVID symptoms and those without. Subsequently, we aggregated these labeled data points to prepare them for machine learning analysis. We employ classification algorithms as we are predicting whether a given set of data points corresponds to either an expectation of severe symptoms or mild symptoms.

We've opted to employ Random Forest and Logistic Regression classifiers for the classification task. Random Forest Classification is a learning method used for classification tasks, it constructs multiple decision trees during the training phase and outputs the class that is the mode of the classes predicted by individual trees. This classification method is widely used in various domains due to its robustness, scalability, and ability to handle complex classification tasks. Logistic Regression is a classification algorithm used for binary classification tasks, predicting the probability that an instance belongs to a particular class.

Following the preparation of the data in the prior phase, we divided it into training and testing sets. We generated models applying the classifier algorithm on the training data and then used the test data set to determine the accuracy of the model.



We have used Accuracy, Precision, Recall and F1 score metrics to judge the accuracy of the model.
The following are the accuracy metrics generated for the models:

1. Random Forest classifier:

Accuracy	Precision	Recall	F1 Score
0.73	1.0	0.077	0.14

- An accuracy score of approximately 73.33% (0.733) suggests that the classifier is correctly predicting positive for severe COVID symptoms around 73.33% of the instances in the dataset. The precision score of 1.0 indicates that when the classifier predicts the positive class, it is always correct. However, the recall of approximately 7.69% (0.077) signifies that the classifier is correctly identifying only about 7.69% of the actual positive instances. The classifier is highly precise in its positive predictions, but it's missing a substantial number of actual positive instances.

2. Logistic regression classifier:

Accuracy	Precision	Recall	F1 Score
0.69	0.4	0.15	0.22

- An accuracy score of approximately 69% (0.69) suggests that the classifier is correctly predicting positive for severe COVID symptoms for around 69% of the instances in the dataset. The precision score of 0.4 indicates that when the classifier predicts the positive class, approximately 40% of the time. However, the recall of approximately 15.4% (0.154) signifies that the classifier is correctly identifying only about 15.4% of the actual positive instances. The F1 score, indicates a trade-off between precision and recall and reflects a moderate overall performance of the model.

We can infer from the above results that both of the models do not have a good overall performance, one of the contributing factor is the limited number of positive labeled data points compared to the negative ones specifically: it is 48 positive instances compared to the 177 negative ones. We have generated synthetic positive instances to check if that could improve the performances, we added 48 instances which increased the number of positive instances to 96 instances. The following are the results:

1. Random Forest classifier:

Accuracy	Precision	Recall	F1 Score
0.85	1.0	0.6	0.75

3. Logistic regression classifier:

Accuracy	Precision	Recall	F1 Score
0.75	0.75	0.45	0.56

The performance improvement is evident when considering all the metrics for both the models, we can conclude from this evidence that it is important to have an even distribution of both the positive and negative labeled datapoints.

VERIFICATION AND VALIDATION OF ALGORITHMIC APPROACHES

For the algorithm-generated biomarker to be used in a real-world clinical setting, it is essential to demonstrate the safety, effectiveness, and performance of the biomarker via a comprehensive clinical assessment. The evaluation process consists of three key components: establishing Valid Clinical Association, conducting Analytical Validation, and completing Clinical Validation.
This paper will center on the analytical validation and clinical validation phases. During the analytical validation phase, we are ensuring that the software is built the right way, meets its requirements, and intended use by gathering and collecting evidence. Moving into clinical validation, we focus on measuring the ability of the software to provide a meaningful output in a real-world health care situation.

ANALYTICAL VALIDATION

During the process of Analytical Validation, our focus shifts to confirming the accuracy, reliability, and precision of the algorithmically generated biomarker. We had followed different steps in the creation of this biomarker: data collection, data cleaning and preparation and finally the application of the algorithm. For each of these steps we adhere to the validation guidelines provided the FDA (General Principles of Software Validation; Final Guidance for Industry and FDA Staff), ensuring compliance with the requisite standards. It is important that we perform analytical validation separately on each of the steps to maintain separation of concern, so that a change in one of the steps does not mean we have to revalidate the preceding steps.
We start with defining the scope of the application of the algorithm which defines the requirements, the intended use and limitations of the software. For instance, for the case study one of the limitations is that the biomarker predicts the severity of the COVID but does not predict the symptoms itself. For the steps mentioned above, we would perform Software Development Lifecycle (SDLC) activities along with validation guidelines so that we ensure that we systematically gather evidence to prove that the software is built the right way.
Regarding data collection, we need to make sure that the measurements collected during the data collection are validated in multiple independent cohorts to ensure its generalizability across different patient populations, ethnicity, age groups, race, gender, and patients with pre-existing disease conditions. We can validate the data cleaning and preparation part by creating test plans which contains test cases to validate each of the steps. These test cases serve to validate the integrity of the data throughout the cleaning and preparation process. Finally, for the validating the algorithmic biomarker, we can use accuracy metrics to understand the performance of the biomarker generated. In our case study, we used accuracy, F1 score, precision and recall metrics to evaluate the performance of the biomarker. We can set a threshold for these metrics so that the biomarker meets a certain benchmark for its performance.

ALGORITHM UPDATE

There are scenarios when the algorithm can be updated or fine-tuned, specifically: updates made to the algorithm parameters and refresh of data post new data collection. It is imperative that we revalidate the algorithm to ensure that the algorithm is working as expected as per the requirements. We would perform regression testing to ensure that the algorithm is working as expected. We would identify and execute test cases which test's the performance of the algorithm, for instance in our case study we had used metrics specifically: F1 score, accuracy, precision, and recall, as we are not changing core structure of the algorithm; updates relate only to the input of the algorithm.

Following the revalidation of the algorithm, we will have to redeploy the algorithm based on the initial location of the deployment. If we are using the cloud infrastructure, single redeployment should suffice. However, when we redeploy it to the devices, we will have to update it on the device by enforcing a software update on the device to ensure the algorithm's implementation is current.

SOFTWARE UPDATE

In cases updates are made to the underlying software or libraries or if a discovered bug is found and fixed, initially we will need to determine the stages which is impacted. These stages encompass data collection, data preparation and application of the algorithm as discussed above. For instance, if the change affects only the core structure of the algorithm, we revalidate this stage which is pertaining to the application of the algorithm by performing regression testing on all the test cases.

CLINICAL VALIDATION

Once the algorithm is technically validated, the biomarker's clinical relevance and predictive value are assessed. This involves testing the biomarker in well-designed clinical studies or trials to correlate its levels or presence with specific clinical outcomes or responses to treatments. Biomarkers clinical validation should be consistent, and results should be reproducible in clinical research settings. The validation of the biomarker should involve comparing it against an alternative method, such as checking heart rate variability through the conventional vital signs check conducted by the Principal Investigator at the clinical site. This comparison allows for a conclusive determination of the target condition and facilitates a comparison against the value predicted by the biomarker. The independent validation should be documented to show the reliability of the algorithmic approach.
Clinical Validation should also encompass testing the algorithm across diverse set of parameters such as diverse demographic patient population (e.g., varying age, ethnicity, race, and sex), as well as population with and without pre-existing conditions like cardiovascular conditions or history of high blood pressure. The outcome of the algorithm should be certified by a clinician and the validation should be compliant as per regulatory requirements. This recommendation would enhance the accuracy of the algorithm.

ALGORITHM UPDATE

There are scenarios when the algorithm can be updated or fine-tuned, specifically: updates made to the algorithm parameters and refresh of data post new data collection. As previously discussed for this specific scenario for Analytical validation, there is no changes to the core structure of the algorithm rather it is changes to the input of the algorithm. This implies that clinically validation is not needed as the clinical relevance and predictive value of the biomarker generated is already established and we focus on ensuring the performance of the algorithm during Analytical validation.

SOFTWARE UPDATE

In cases updates are made to the underlying software or libraries or if a discovered bug is found and fixed, we will have to determine and assess the severity of the changes. Depending on the severity of the change, we will have to

decide if the validation needs to be performed again. For instance, in the case of software update, if there is a version upgrade in some of the package utilized, this can be classified as a low risk change as there is no change in core structure of the algorithm and its performance is ensured during analytical validation; clinical validation is not necessary.
Conversely, if a new algorithm is introduced to create the biomarker, this scenario will be classified as high risk as this algorithm was not validated previously in the clinical context, validation would be required again. The same principle applies for the scenario's where the bug is discovered and fixed, we will assess the severity of the bug to determine if revalidation is necessary.
We can employ the International Medical Device Regulators Forum (IMDRF) risk categorization principles, FDA's benefit-risk framework as detailed in the FDA paper (Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)) to categorize risk in the context of the change being made to the software. The framework considers the significance of the information provided by the to the healthcare decision along with the identified state of healthcare conditions or situations.
In the situation, when a change happens in the middle of a clinical validation study for the software depending on the category of risk the change belongs to, we will have to document the change and report it to the FDA if it falls under high category risk. This ensures transparency regarding the software changes, and keeps the FDA well-informed about the changes.

CONCLUSION

We have devised a case study where we created a predictive biomarker by application of software and technology on the Welltory COVID-19 dataset to predict the severity of the symptoms for a patient testing positive for COVID-19. The purpose of the case study was to understand the steps involved so that we could provide recommendations to verify and validate the biomarker; so that it can be used in clinical trial. Recognizing the pivotal nature of data collection from wearable devices; we noticed that the performance of the biomarker improved when there was even distribution of the positive and negative labeled data points. Furthermore, we propose that the biomarker is technically validated against measurements collected from multiple independent cohorts to ensure its generalizability across diverse patient populations. We recommend using the accuracy metrics as demonstrated in our case study to technically validate the biomarker; by setting a threshold hold values for these metrics enables us to ensure a performance benchmark; the values can be determined based on the use case and their severity. Upon successful technical validation, we recommend clinical validation via well designed clinical studies or trials; with the aim of comparing the biomarker against the determination of the target condition made by the Principal Investigator; ensuring its reliability and accuracy in real-world healthcare settings. We also explore how algorithm and software changes impacts the analytical and clinical validation process of the biomarker.

REFERENCES

1.Welltory COVID-19 dataset. Retrieved from https://github.com/Welltory/hrv-covid19

2.FDA. (2022). General Principles of Software Validation; Final Guidance for Industry and FDA Staff. Retrieved from https://www.fda.gov/media/73141/download?attachment

3.FDA. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). Retrieved from https://www.fda.gov/media/122535/download?attachment

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Kartik Shetty
Company: Takeda
Address: 650 Kendall St, Cambridge, MA 02142
Email: kartik.shetty@takeda.com