Artificial Intelligence for risk prediction and/or stratification

Bipasa Biswas
CDRH, FDA, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract
Artificial Intelligence (AI), is a technology that uses algorithm and software to combine large amount of data to learn automatically from patterns or features in the data and interpret underlined complex phenomena. AI is currently at the center of the medical horizon, expected to be used on an ongoing basis to change care pathways by expediting early detection and improve patient access to needed healthcare. Diagnostic devices utilizing Artificial Intelligence (AI), Deep Learning (DL) or machine learning (ML) often generate a risk score and/or probability of an outcome/event. This presentation will give an overview of the AI, DL and ML and discuss issues and challenges with probability scores.

Key Words: Artificial Intelligence, probability score, diagnostic devices.

1. Introduction

Biomedical devices are seeing more and more the utilization of artificial intelligence (AI) due to the availability of large data and computation power. Biomedical imaging devices have led to terabytes of data (pixels or voxels from images). The imaging devices non-invasively explore inside the human body before complex procedures and have seen utilization of AI for three broad categories-1) image segmentation (methods to distinguish between biologically relevant structures such as tissues, organs and pathologies); 2) image registration (aligning images); and 3) image based physiological modeling. Examples of such sources of images being X-ray, Computed tomography (CT), magnetic resonance imaging (MRI), Ultrasound Imaging, single-photon emission computed tomography (SPECT), positron emission tomography (PET), and visible light imaging. Examples in in-vitro devices include digital pathology, images of blood cells.

1.1 Artificial intelligence, Machine learning, Deep learning

In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals. Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving". Artificial Intelligence has been broadly defined as the science and engineering of making intelligent machines, especially intelligent computer programs (McCarthy, 2007). Artificial intelligence can use

different techniques, including models based on statistical analysis of data, expert systems that primarily rely on if-then statements, and machine learning.

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine Learning is an artificial intelligence technique that can be used to design and train software algorithms to learn from and act on data. Software developers can use machine learning to create an algorithm that is 'locked' so that its function does not change, or 'adaptive' so its behavior can change over time based on new data.

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised.

A Neural network is highly parametrized model, inspired by architecture of the human brain, -a machine that with enough data could learn any smooth predictive relationship.

2. Prediction

Artificial intelligence and machine learning techniques are often used for predictions in biomedical devices. Prediction present two problems; firstly, construction of an effective prediction rule and secondly to estimate the accuracy of its predictions. Construction of prediction rule constitutes of development phase of the model on a training data set. Once the model is constructed on a training data set, it is important to address the prediction error for a new case obtained independently of the training data set. They are addressed via internal validation where estimation of accuracy, or rather prediction error, is either addressed using model-based approaches like Mallow's Cp estimate and Akaike information criteria (AIC) or using cross-validation approaches. There are several approaches to internal cross validation discussed in literature like random splitting, leave one out, J-fold validation, temporal and spatial split of the data. Just as estimating the prediction error using data from training set usually underestimates the prediction error, a random split generates an optimistic prediction errors due to similarity in the data set between training and testing. Altman and Royston (2000) refers to it as the weakest validation. A clinical validation of the prediction is best assessed using an external data set separate and independent from training data set and after any internal validation. And the external validation is necessary to establish that it works satisfactorily on patients other than those from whose data it was derived.

The prediction in medical devices could be for diagnostic (probability that a certain disease or condition is present) or prognostic (probability that a specific event will occur in future). It can serve two purposes, one at group level: stratification i.e. classification into risk categories (to inform on treatment decisions or stratify patients by disease severity for clinical trials) and the other at individual level: risk probability i.e. probability of an event (e.g. disease). Fundamentally, the two are different as an excellent model may successfully distinguish between high and low risk patients but ability to provide informative prediction at the individual level, such as the patient's expected survival time with a narrow 95% confidence interval, is almost always limited.

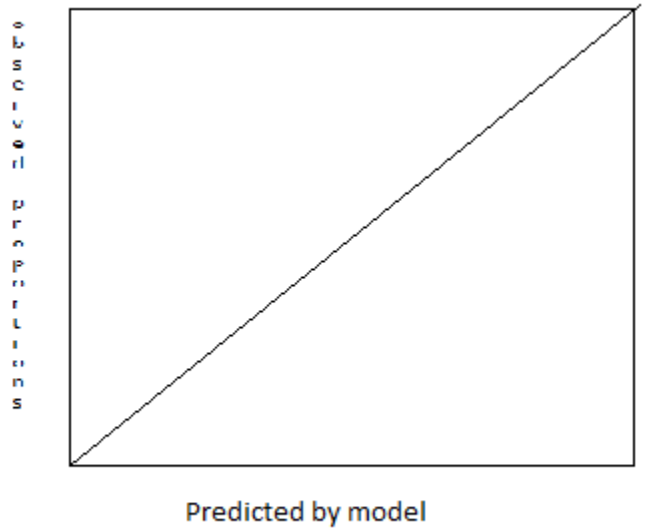
2.2 Design

The design of such studies to build as well as validate the models need to take the context of use into consideration. In particular, the data set from the patients on whom it is to be used in practice (both the end users of the models and the patients/subjects). The size of the dataset plays a role as smaller data set may result in unstable predictors. Further, it is important to define the types of outcomes (or events) and particular attention needs to be paid to assess the number of events per variable (a general rule prescribed in literature is about 10-20 events per predictor variable). In order to address accuracy, it is important to prespecify a clinical reference standard, which is independent of the predictors or the outcomes of the prediction model and is blinded from the model outcome.

2.3 Model Validation

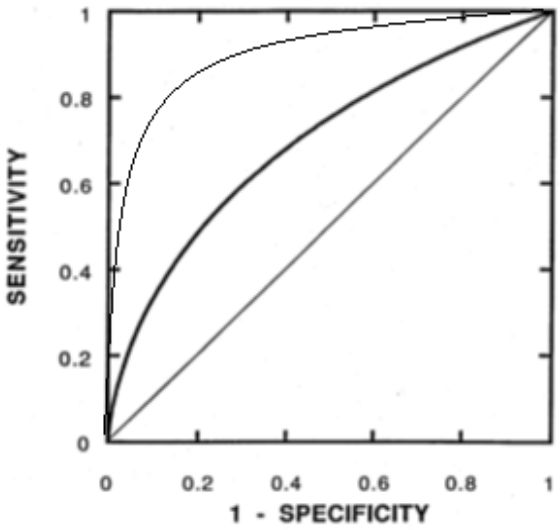
Steyerberg and Vergouwe (2014) address seven steps to model development and validation. This paper focused on the stages of model validation. Essential parts of prediction models is assessed by internal as well as external validation. The need for internal validation as addressed by Steyerberg and Harrell is stressed in their commentary. The validation is addressed via calibration and discrimination and the usefulness of the model in the clinical context is discussed from decision analytic framework.

Calibration addresses the agreement between observed endpoints and predicted values. Often and commonly used is, the Hosmer-Lemeshow goodness of fit test. However, the test has its limitations as it fails to indicate the direction of miscalibration, and it is based on p-value which depends on sample size. Further, the test utilizes grouping of patients (usually deciles) which are arbitrary. Steyerberg (2014) proposes assessment of calibration via plots (predicted versus observed) which is characterized by deviation from the y=x line (Slope ‘1’ and intercept ‘0’).



Discrimination is the ability to distinguish patients with the endpoints (events) from patients without the endpoints and is assessed via c-statistics (AUC of ROC curve). AUC

is the area under the Receiver Operating Characteristics (ROC) curve. The closer the value of AUC to 1 the better the discrimination, while, AUC of 0.5 indicates poor discrimination



Calibration and discrimination, although important aspects of prediction models, do not assess clinical usefulness. The decision thresholds for clinical decision -classify patients (severity of disease or low risk and high risk) utilize decision analysis as discussed by the authors.

2.4 Importance of external validation

While internal validation addresses the model reproducibility, generalizability or transportability is addressed via external validation on a data set of patients completely separate and independent from that used to train and develop the model. The validation data sets are addressed as three different forms: study patients who were more recently treated (temporal validation), from other hospitals (geographic validation), or treated in fully different settings (strong external validation). A strong external validation is the best form of validation to assess the generalizability of the model.

The external validation (also referred as clinical validation in the context of medical device evaluation) is assessed on a separate and independent data set of patients, from the intended use population, after fixing the model and any cutoffs.

3. Summary

AI/ML/DL are being used more and more due to availability of data (use in imaging data). Often AI/ML/DL are used to generate predictions for biomedical use. In the development as well as validation, it is important to understand the context of use e.g. the population on whom it is to be used, the endpoints, clinical usefulness etc. The prediction validation involves an internal validation to assess reproducibility and an external validation to assess transportability to the intended population. There are several approaches to an internal validation like leave-one-out cross validation, J-fold cross validation etc. The internal validation should not be confused with an external validation and that to address the usefulness it is important to address external

validation. The model and any thresholds need to be fixed before the external clinical validation.
Just like for diagnostic test evaluation there is STARD, there is a consensus standard for transparent reporting called TRIPOD to address the appropriate reporting for prediction models,

Acknowledgements

The author acknowledges help and support of Dr. Changhong Song.

References

1. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
2. McCarthy, J. (2007). *What Is Artificial Intelligence?* Stanford University, Stanford, CA. Retrieved from http://jmc.stanford.edu/articles/whatisai/whatisai.pdfExternal
3. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statist. Med.* 2000; 19:453-473.
4. Efron B, Hastie T. Computer Age Statistical Inference. Cambridge University Press (2016).
5. Bleeker SE, Moll HA, Steyerberg EW et al. External validation is necessary in prediction research: A clinical example. Journal of Clinical Epidemiology 56 (2003) 826–832
6. Janes H, Pepe M, Gu W. Assessing the value of risk prediction by risk stratification. *Ann Intern Med.* 2008;149:751-760.
7. Moons KGM, Kengne AP, Woodward M et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart (2012). doi:10.1136/heartjnl-2011-301246.
8. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart (2012). doi:10.1136/heartjnl-2011-301247.
9. Steyerberg EW, Vickers AJ, Cook NR et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010 January ; 21(1): 128–138. doi:10.1097/EDE.0b013e3181c30fb2
10. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal (2014) 35, 1925–1931.
11. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology 69 (2016) 245e247.
12. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162:55-63. doi:10.7326/M14-0697.