Issues with Training, Testing and Validation Datasets in the Development of Diagnostics Devices

R. Lakshmi Vishnuvajjala
Food and Drug Administration, Center for Devices and Radiological Health,
Silver Spring MD 20093

Abstract

Model development and validation are critical parts in the development of classifiers, or diagnostics devices as they are called in submissions to FDA. The integrity of methods used to develop and validate classification models ensures the performance of the diagnostics devices. There is a lot of confusion about training, testing and validation datasets, as well as internal and external validation of models. We will discuss some good practices for developing and validating classification models in diagnostic devices. Specially, we will investigate some problems frequently encountered with training and validation datasets which can lead to overly optimistic estimates of performance metrics.

Key Words: Validation, Testing, Training, Classifier, Diagnostic Device

Diagnostic Devices are regulated by the Center for Devices and Radiological Health at the Food Drug Administration. A medical device is an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is recognized in the official National Formulary, or the United States Pharmacopoeia, or any supplement to them, intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals, or intended to affect the structure or any function of the body of man or other animals, and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of any of its primary intended purposes.

Diagnostic Devices are all tests/classifiers which classify patients into different groups. A model, or a classifier, separates patients into classes with or without the condition, or into different grades based on relevant parameters. It could be a prognostic marker which separates patients into groups with different outcomes under standard of care. It could be a predictive marker that separates patients into groups with different outcomes when given a specific treatment. It could separate patients into different groups with respect to risk of disease, e.g., a particular mutation of a gene.

The reason for developing classifiers is to allow reliable classification of future patients. It is not enough to show that it predicts outcome in the data used to develop it; it needs to perform well for other groups of patients in the intended use population. Building and testing a classifier usually involves thee steps. Building the classifier; fine tuning it to fit the study objectives; and then validating on different, independent data to assure that the results can be extended to the intended use population.

Training dataset is the data on which the test or classifier is developed. Internal validation, which is sometimes done on all or part of the training dataset, is used to fine

tune the model. External validation on a separate independent dataset is generally referred to as testing. During the development of the classifier, one needs to be blinded to the test data. The classifier needs to be finalized and fixed (locked down), before one can look at the testing dataset.

In submissions to CDRH, we have encountered issues with the data used in different stages of building and validating the classifiers. Some of the issues are (i) assuming internal validation is external validation, or testing, when it is part of training; (ii) looking at the test data before the classifier is fixed; using data that has been previously used to support a different/ modified classifier, and (iii) using parts of the same dataset to build and validate the classifier, and splitting the dataset randomly into the two parts.

For modeling and internal validation, it is fairly common to use different parts of the same dataset. One part to build the model, and the other to fine tune, or internally validate the model; or use all the data to build the model and use bootstrapping or k-fold cross validation to validate. Both of them however, are internal validation, and cannot substitute for external validation, which should be done on an independent dataset to which people developing the model should be blinded until the model is fixed. There have been several cases where internal validation was assumed to be external validation, or testing.

When a model was developed and validated, in some instances, people were going back to modify the model based on new information. One example could be a test that classifies people into two groups; patients above, or below a specific level for the marker. Two years later, it was felt that the patients should be classified into three groups rather than 2. The same data that was used for external validation in the first case was also used for external validation in the second case. But this data was known as it was used previously, and the three regions can therefore be chosen to optimize, or fine tune the performance. Therefore, it is acceptable for model building and internal validation, but definitely not for external validation. In fact, even for internal validation, it can be overly optimistic (Altman et al, 2009).

Sometimes, a dataset is split into two parts to develop the model and for external validation. This can happen when a classifier fulfills an unmet public health need and a retrospective dataset that meets all the requirements is available. Particularly true when long follow-up data is needed since new data cannot be obtained quickly. Even then, if the available dataset is large enough to get two samples from it, it is preferable. But when it is only big enough to split into two parts, that is done. And more often than not, it is split randomly into two parts. Randomization which is considered good for most things is not good in this case, for exactly the same reasons. This is different from drawing a random sample from a large population and more like assigning patients to two treatment arms. We randomize patients in a clinical trial as randomization balances all covariates, known and unknown. Similarly, a random split creates two very similar groups. So, training on one and testing on the other is very similar to testing on the training set, and overestimates performance. So, splitting in some other way, chronologically for example, or by site, is better. Since the model, or test, will be used on different patients, it is better to get validation data which is totally independent. But in the event we need to develop and test a classifier quickly and a retrospective dataset is available, we run into this issue. This is less than ideal, but accepted sometimes with enough safeguards, just as we accept historical data.

Gary Collins et al (2014) describe a review of articles on multivariable prediction from 2010. They conclude that a vast majority of studies describing some form of external validation were poorly reported, and state that this is consistent with other reviews of prediction models. They propose an initiative, called Transparent Reporting of a multivariate model for Individual Prognosis Or Diagnosis (TRIPOD). These are essentially guidelines for reporting, rather than developing and validating the model, but planning for appropriate reporting should also lead to an appropriate study design.

The TRIPOPD initiative was started by developing an extensive list of items based on a review of the literature which was reduced after a web-based survey and revised during a three day meeting in June 2011 in Oxford, UK, with methodologists, health care professionals, and journal editors. The list was refined during several meetings and email discussions and resulted in a checklist of 22 items, deemed essential for transparent reporting of prediction model studies. The TRIPOPD statement was published in Annals of Internal Medicine in January 2015, and is also published in ten other journals.

In Summary, the point of developing a classifier or model is to be able to apply it to new patients. This would mean different centers and times, and possibly different countries. The data to test, or externally validate a model or classifier should be independent of the data used to build or internally validate the model or classifier. Splitting one dataset into two randomly for developing and external validation overestimates the performance of the classifier. It is not advisable even for internal validation as it increases the chance of false discovery. We want the modeling and validation datasets to come from the same intended use population. But we don't want to validate on the same data on which we trained, or built the classifier. Training and testing on datasets that are derived by a random split from one dataset is the next closest thing, and would overestimate its performance in the intended use population. In addition, people involved in building the model or classifier should be blinded to the data used for external validation.

References

1. Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman and Karel G. Moons. Transparent Reporting of multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Annals of Internal Medicine 2015; vol 162: 55 – 64.

2. Gary S Collins, Joris A de Groot, Susan Dutton, Omar Omar, Milensu Shanyinde, Abdelouahid Tajar, Merryn Voysey, Rose Warton, Ly-Mee Yu, Karel G Moons and Douglas Altman: External Validation of multivariable prediction models: a systematic review of methodological conduct and reporting**.** BMC Medical Research Methodology 2014, 14:40.

3. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: Validating a prognostic model. BMJ 2009;338: 1432-1435.

4. Royston P, Moon KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009;338: b604.

5. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009; 338: b606.

6. Altman DG, Royston P. What do we mean by validating a prognostic model. Statistics in medicine 2000; 19: 453-473.