Validation Methods to Promote Real-world Applicability of Machine Learning in Medicine

Riyad B. Rafiq
Department of Computer Science and Engineering,
University of North Texas, Denton, TX, USA

Francois Modave
Department of Health Outcomes and Biomedical
Informatics, University of Florida, USA

Shion Guha
Department of Computer Science, Marquette University,
Milwaukee, WI, USA

Mark V. Albert
Department of Computer Science and Engineering,
University of North Texas, Denton, TX, USA

ABSTRACT

The impact of Artificial Intelligence (AI) on health care has been dramatic; however, there is a considerable degree of skepticism among clinicians about the real-world applicability of advanced predictive models; for this reason, it is particularly important to emphasize the need for proper model validation in machine learning. Often model skepticism is well-placed as modelers may overclaim the real-world replicability for their models, understate the known limitations, or simply not be aware of the hidden limits of the modeling approach. Educational approaches limited to rigorous and thorough justification of all model design decisions may not be practical given model complexity. This also becomes more challenging as state-of-the-art models with the highest benchmark accuracy are becoming less interpretable, e.g. ensemble methods or deep learning. However, in the same way that test-driven development has been a successful paradigm to navigate the complex coding landscape through a focus on testable results, we have observed a similar improvement in modeling strategy when the focus of a predictive model is driven by validation targets rather than more abstract, theoretical concerns. In this study, we provide an overview of the common limitations of model validation methods in medicine. We then present solutions to address such limitations, with a focus on strengthening the validity of predictive models.

CCS CONCEPTS

• **Applied computing Health informatics**;

KEYWORDS

Validation methods, Natural variability, Evaluation metrics, Precision medicine

ACM Reference Format:
Riyad B. Rafiq, Francois Modave, Shion Guha, and Mark V. Albert. 2020. Validation Methods to Promote Real-world Applicability of Machine Learning in Medicine. In *2020 3rd International Conference on Digital Medicine and Image Processing (DMIP '20), November 06–09, 2020, Kyoto, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3441369.3441372

1 INTRODUCTION

Machine learning and AI are used extensively in medicine and further leverage the massive amount of data collected to improve clinical decision making [1], [2]. A major impediment in the medical application of machine learning and AI is establishing the reliability of these predictive models in practice [3], [4]. There are many scientific and social factors favoring the publication of predictive models with inflated accuracy, and fewer incentives for machine learning practitioners to take the additional steps to better evaluate the real-world utility of their models [5]. Clinicians alone have little time or resources to vet data driven models on an individual basis, and when models are applied automatically, e.g. in electronic health record (EHR) systems, the skepticism and improper validation during the vetting process can delay proper application of beneficial predictive models. Additionally, clinician judgement is often based on reportable decision-making steps or developed intuitions that have been verified as accurate over the course of a career, while such interpretability is often not available to machine learning models without significant additional steps [6], [7]. When relatively opaque machine learning models consistently underperform in clinical settings in relation to published performance metrics, it is not unreasonable to treat such models with skepticism. Therefore, it is critical to teach the proper strategies for developing models for health care [8].

Predictive models often augment standard clinical practice by performing a preliminary screening based on routinely collected data. For example, in cardiovascular medicine the most promising directions for AI models include automated risk prediction which can alert a clinician to a person at high risk of a cardiovascular event based on a number of factors that a practicing physician may not have access to in a timely way [9], [10]. This can also include the prediction of major complications, including death, following surgery [11]. In fact, available public data sets enable efficient creation and testing of models to predict a variety of diseases including heart disease, diabetes, liver disease, dengue fever, and hepatitis [12]. Data-driven screening helps to minimize the risks of classifier mistakes, but improved models can help make the process more reliable.

We also emphasize the need for assurances through properly validated models as a means of eventually moving beyond the requirement for immediate human interpretability of decision making.

DMIP '20, November 06–09, 2020, Kyoto, Japan
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8904-4/20/11…$15.00
https://doi.org/10.1145/3441369.3441372

Much of machine learning policy has focused on the need for model decisions to be interpretable [13]. This need for accountability has resulted in the EU declaring that users may have a "right to an explanation" from machine learning models applied to their data [14]. Researchers who create methods of explaining the results of machine learning note the limitations [15], [16]. In particular, requiring language-level, simple mathematical, or visual descriptions of decision-making steps may be unnecessarily limiting [17]. The practical utility of knowing what led to a diagnosis is beneficial, however, there are contexts and modeling strategies where a requirement for thorough understanding of the decision making would limit the accuracy a model could achieve [18]. An emphasis on proper validation rather than detailed human language-level understanding is particularly beneficial in the modern era as state-of-the-art models in terms of performance are often the most challenging to explain in a classical way.

We have structured this paper to draw attention to several common mistakes in establishing validity of machine learning models in medicine; the most common are presented in Section 2, and methodological improvements are addressed alongside the mistakes as they are presented. The later sections are structured to provide the necessary tools to address the more sophisticated concerns. Section 3 addresses one of the main limitations of reported machine learning-based models for clinical application - the lack of proper sampling for training and testing. Section 4 addresses why it is important to use the right metrics for validating a predictive model in medical contexts. This section emphasizes classification and includes a discussion of regression metrics for judgements of intensity or clinically useful scoring. Finally, Section 5 presents how to customize models for any specific application with explicit suggestions and illustrative examples. An overview of using proper validation of machine learning approaches in medicine is provided in the conclusion.

2 COMMON VALIDATION MISTAKES

We list the following common mistakes in assessing machine learning model prediction when applied to medical applications:

- Equating validation set and test set: Only a simple cross-validation reported

Some computer scientists are aware of the need to separate training and test data and come to the conclusion that any use of cross-validation satisfies this concern. Others are aware of the issue of overfitting even with cross-validation, but understand the nuance that as long as a limited number of model variants are tested, overfitting through cross-validation is minimized. This may have been a viable strategy in the past when fewer model variations were tested in each study and the use of reporting k-fold cross-validation results alone would be less likely to produce biased metrics. However, now modern hardware enables the testing of many thousands or even millions of model variations through feature selection, hyperparameter tuning, model selection, or even ensemble methods. The potential for overfitting during a simple cross-validation procedure is much more likely and is a growing concern, with a number of ways to address it [19], [20], [21].

- Shared variability of train and test samples

For proper model validation in machine learning, the test set should be independent of the training and validation sets to avoid inflating reported model accuracy. However, contamination can occur between training and test sets through a variety of mechanisms that may not be readily apparent to the model designer or evaluator.

Many models vary depending on the place in which they are applied. Clearly there is a need to test for different subjects, but dependence can also come from the same experimenter performing a procedure, or the same clinic, or the location in which the population of people reside. In short, any variability present in the population to which a model will be applied should be systematically considered when evaluating a given test set. Modelers may prefer to avoid such variability for cleaner interpretability, but in clinical practice this variability is the inherent nature of the problem and a common source of disconnect between reported experimental accuracy and the accuracy observed in real-life application.

- Relying on the default choice of evaluation metrics

Accuracy is the classification metric that is easiest to understand but also the one most open for abuse. As a classic example, many naive threat detection systems can easily be 99.9% accurate by labeling all cases as non-threatening. Using recall, precision, or their combination as an F1 score, helps to address this concern, though other interpretable solutions will be discussed later.

For regression, there is an emphasis on mean squared error (MSE) because of its prominence as the cost function of choice in introductory statistics courses and a first exposure to regression. This reliance on MSE can lead to poor models in practice, particularly for models addressing a wide range in target values. The right choice of metric involves judgements on how much the model should address outliers compared to the more common cases, and also what types of errors would be considered equally costly.

- Lack of validation in software industry

Though such a mistake is not common in the biomedical community, this is not an unusual observation in papers focused on software development. For example, in a study by Zelkowitz and Wallace of 600 articles in software engineering, it was found that 30% of the papers had no validation when it could reasonably be expected. Additionally, 34% of the papers were validated, but only by data collected from the authors themselves [22]. (Figure 1)

A lack of independent validation may be from naive modelers who are not aware of basic overfitting concerns, but more often this occurs due to a lack of emphasis by the designers. Medical device prototype articles that focus on system design and implementation may fit a predictive model then perform minimal testing assuming that is left for future work. In other cases, efforts may be made to address reviewers to demonstrate efficacy with only an analysis of their own data or very limited convenience samples, with little concern for proper statistical inference.

- Non-representative samples: Absence of natural variability in population

This error occurs when the population the model will be applied to is not well represented in the data sampled. This can range from common concerns like demographics not being represented to more nuanced concerns that can impact model performance,

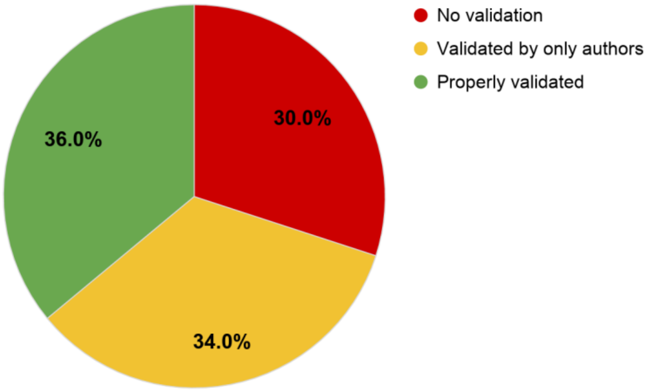

Figure 1: Survey of validation strategies in software engineering by Zelkowitz and Wallace [22]. 600 published software engineering articles were evaluated on the method of validation used, designating "no validation" or "validation only by authors" in articles where additional validation could be reasonably expected.

such as fitting a model for particular subpopulations in particular contexts, e.g. patients with a particular disease severity at home versus in the clinic.

There is a general tradeoff in acquiring data with minimal variability to remove extraneous factors versus more natural data which may lead to quantitatively poorer models but more robust performance outside of research settings; concerns of realistic application, when addressed, may be left for the discussion section. However, to evaluate clinical significance it is important that natural variability is present in the test cases; also, as discussed in later sections, such natural variability should be present during training and validation to improve performance in practice.

3 LIMIT THE IMPACT OF APPLYING MANY MODELS DURING CROSS-VALIDATION

Simply identifying the common mistakes made in model validation is a helpful first step. One of the most prevalent issues is the lack of proper testing of cross-validation results [23], so in this section, we briefly step through some common ways in which this limitation and others previously mentioned can be properly addressed. We end this section with a strong recommendation to consider nested cross-validation to avoid concerns with arbitrarily selecting a hold-out test set.

3.1 Pre-registration

Pre-registration is becoming more common in science [24]. This allows reviewers to note how many models and statistical tests are applied when significant results are reported to judge the potential for false positive results. Tools include pre-registration at clinicaltrials.gov and the center for open science OSF study pre-registration. When pre-registration is not feasible, an exhaustive reporting of model variations applied to any portion of the data, but not fully elaborated upon in the manuscript, can help provide similar information for judgements about overfitting due to multiple testing.

3.2 Limit Variations Tested during Cross-validation

Cross-validation is a valuable tool for finding the best model parameters, however overfitting is a common problem particularly when many model variants are tested [25]. The impact of multiple testing during cross-validation is minimized if only a few models or hyperparameters are tested. Studies which do not use a hold-out test set during cross-validation rely on this fact, knowingly or not. If the model performance is to be evaluated using cross-validation but without a separate test set, the potential for overfitting should be documented in the manuscript. Furthermore, if it is observed that many different folds during validation testing select the same set of parameters producing the preferred model there is stronger evidence that the model is not overfitting but simply picking the best model variant; demonstrating the variation in models selected for each fold can also provide additional evidence that overfitting is not occurring.

3.3 Nested Cross-validation

Even though simple training/test cross-validation alone is not sufficient to completely remove the potential for overfitting [26], more sophisticated forms of cross-validation do avoid overfitting. In the classic training/validation/test set paradigm, the validation set can be used to select all aspects of the model and a hold-out test set is used for the final evaluation. However, a hold-out test set may not be recommended when there is insufficient data. A hold-out test set which is too small may not represent the population well, and a larger test set may remove a significant amount of data that may impact performance during training and validation iterations. If there is limited data for a hold-out test set, but many model variations will be tested likely leading to overfitting, nested cross-validation should be considered [27].

Like classic cross-validation, during nested cross-validation the data is split into "outer" training and testing partitions until all the data has been a part of the testing partition; however, each training partition is further divided using an "inner" cross-validation loop consisting of training and validation partitions. In this way, multiple hyperparameters, models, or other model selection mechanisms can be used to select the best model for application to the test partition for each round.

4 VALIDATION METRICS APPLIED IN MEDICINE

A model can be optimally trained and properly evaluated when the metric of success best matches the needs of the anticipated environment the model will be deployed in. In this section, we step through a series of metrics that we believe should be considered to improve clinical applicability of machine learning models. Each section will begin with common metrics and end with metrics in which greater adoption could improve clinical model applicability.

4.1 Classification Metrics

4.1.1 Metrics for Binary and Multi-class Classification Medical Problem. In the classic case of disease screening and diagnosis, the use of sensitivity and specificity, rather than simply quoting an overall accuracy metric, is well known [28]. The use of one of these two metrics without fixing or addressing the other is generally not acceptable as one can often arbitrarily be increased by minimizing the other. This is standard practice for binary classification problems in medicine. There is an increasing prevalence of multiclass classification problems in medicine [29], [30], [31], [32], it may be beneficial to emphasize the multiclass counterparts of sensitivity and specificity. Recall is the percentage of correctly classified samples over all the samples of a given class. Succinctly, Sensitivity is recall for the positive group while specificity is recall for the negative group. The complement to recall is precision. Precision is the percentage of correctly classified samples over all the samples of a given class as identified by the classifier. Whereas recall is an intuitive and readily interpretable metric for evaluating how well a classifier will perform for an item of a known class, precision provides the equivalent reliability metric for a sample as identified by the classifier.

4.1.2 F1 Score to Balance Precision and Recall. When the application of the multiclass classifier does not readily distinguish which metric is preferred, the harmonic mean of precision and recall, F1 score can be used for a given class. The practical advantage of the harmonic mean is that it heavily penalizes F1 score when either precision or recall for a class is low. For example, let us consider a naive binary classifier that always indicated a positive case for an event which is rarely positive, say only 2% of the time. This would result in 100% for recall for the positive case (sensitivity), however, the precision for such a classifier, if samples are chosen randomly from the population, would be only 2%. The arithmetic mean of precision and recall in this case would be 51% which would seem to give too much credit to this terrible classifier. However, the harmonic mean, for which the F1 score is based, would report only $\sim$ 4%. Precision, recall, or F1 score can be selected for optimization in a straightforward way, especially for model selection using cross-validation. Modelers would benefit from being aware of these options in their model package of choice.

4.1.3 Context of Misclassification by Confusion Matrix. For model tuning, particularly with selection during cross-validation, an overall metric is necessary to select the best model. However, for interpretation and human evaluation it should be noted that, in most cases, overreliance on a single metric for a multiclass classification is an oversimplification of the results, as some classes are more readily distinguished than others, and the manner in which misclassification occurs provides useful information. In this case a provided confusion matrix or reported scores for each class and common misclassifications are more valuable to provide context.

4.1.4 Addressing Class Imbalance. One of the main concerns in providing an overall classification metric is how class imbalance should be addressed. There are a few common options in averaging scores over multiple classes to address this:

- Precomputed: Determine the results across classes and calculate the true positives, false negatives, and false positives to compute overall metrics. This is a common default when optimizing based on overall accuracy.
- Unweighted average: Average the score reported for each class and calculate the average, regardless of the number of samples in each class.
- Weighted average: Average the score across classes, but weigh the contribution of each class by a function of the number of samples. If single errors from each class are to be approximately equal, the weight can be proportional to the size of the class, however, arbitrary weights can be used if certain classes are more critical, or costs for particular errors are known.

4.1.5 ROC-AUC Curve for Evaluating Model Efficiently in Binary Classification. In the case of binary classification, a model can be fit with an option to tradeoff between higher sensitivity or higher specificity. If a reasonable range for either of these parameters cannot be estimated, one may evaluate the overall discriminability of the model over a range of sensitivities and specificities.

The ROC curve (Receiver Operating Characteristic) is a plot which displays the change in the sensitivity at fixed specifiers. The common terminology is the true positive rate (equivalent to sensitivity) is plotted along with false negative rate (1 - specificity). ROC curves are used in medicine for the evaluation of a clinical test. In order to distill the information in the ROC curve to a single metric to represent discriminability, the area under the ROC curve is often used - abbreviated ROC-AUC. A naive classifier would result in a 0.5 ROC-AUC while a perfect discriminator would evaluate to 1. This is a valuable metric in evaluating binary classifiers when the application domain is not clear about an appropriate specificity for direct comparison of different classification strategies.

4.2 Regression Metrics

The evaluation metrics used when predicting a numeric value can often be more challenging to interpret than for classification. Generally, the goal is to minimize a collective error metric over a set of test cases. There are a variety of ways to combine those errors and each method is based on a set of assumptions of which the modeler should be aware.

- Minimizing mean square error (MSE) is a common default, but comes with assumptions: For most people this is the default scoring metric used. There are many reasons for this, with the primary reason being that a first exposure to regression is often with linear regression using least-squares techniques [33]. Mathematically, summed squared distances represent natural Euclidean distance and minimization techniques for quadratic cost functions are fairly simple. Practically, if there is no reason for a person to prefer any of the metrics below, this is not an unreasonable choice as it penalizes large errors more than a series of smaller errors, but not excessively so. In some cases, modelers report the sum of squares error, however, to aid in interpretability the root sum of squares error (RMSE) should be reported as the units will be the same as the original measurement and readily interpretable.
- Minimizing mean absolute error (MAE) for weighing errors across all samples more: If the goal is to strictly minimize the sum of errors between regression estimates and the true values, the absolute error metric is the preferred metric as it penalizes errors in proportion to their magnitude (rather than the square in MSE). Though sum of squares minimization is more common across regression software packages, often an equivalent routine exists to minimize absolute error. This metric is also simple to select for any model during cross-validation during model selection. Note, however, that both MSE and MAE cause trained models to bias toward correction of high magnitude values [34].
- Minimizing log error to penalize percent error: If the goal is to minimize the percent error rather than the magnitude of the error directly the metric of choice is the log error. This is a result of the natural property of the logarithm as multiplicative errors become additive errors when a logarithmic function is applied. Note that one can either formulate the log error with absolute error terms or as a sum of squares. In the case of the log of square errors, the preferred metric is the root mean squared logarithmic error (RMSLE) which penalizes similar percentage errors but penalizes higher percentage outliers to a greater degree than absolute log error would.

4.3 Explicit Cost Function with Bayesian Decision Theory

The previous metrics all weigh errors differently depending on the type of error, however, all do so in relative terms without explicitly referencing the cost of an error. In medicine, although clinicians often weigh competing risks and costs on a daily basis there is a natural aversion to making those costs explicit. In any case, the tools of Bayesian decision theory can be applied in scenarios with explicit a priori costs or implicit costs to be estimated from repeated decisions. If the costs of an error are known or approximated, Bayesian decision theory can be used to build a model which explicitly minimizes such costs. If the task is a multi-class classification, a matrix can simply be filled in with the relative costs of each type of misclassification. For example, errors in screening tests and diagnostic tests can often be weighted differently depending on the consequences.

5 TAILORING MODEL TO THE INTENDED APPLICATION

5.1 The Case for Collecting More Realistic Data

Predictive models are often developed with specific populations and contexts. This may lead to developed models that perform better for some populations in particular contexts while being rendered inapplicable to others. Validity can be lost when any system is used for a population or in a context for which it wasn't properly tested. The goal of the reviewer, funder, and ultimately consumer of medical predictive models is to assure the test scenarios reflect the target population, and the proper context for that population. This is often summarized as seeking more natural variability in the model rather than limited, lab-acquired data sets.

When more realistic data is collected, the goals of the model have to adapt to the challenges in the research approach [35]. More natural data collection will likely be more difficult depending on the way in which the data is acquired (e.g. from patients, from hospitals, by natural observation) and accuracies may suffer when there is insufficient data as a result of these difficulties. With more variability, more data is required to reach peak performance. A recommended way to assess the need for more data is through the use of a learning curve which plots the accuracy of generated models in the test sets relative to the amount of data collected. With difficult-to-collect data (e.g. rare clinical patient samples) an upward learning curve can help support the potential for a modeling approach with more data when there are practical limits to preliminary real-world sampling.

5.2 Emphasizing Subject-wise Cross-validation

Many researchers create predictive models with an understanding that there is more variability across individuals than within individuals. When the goal of a predictive model is to predict future status of an individual using prior data collected on that individual, one can use validation methods that split an individual's data between training and testing. When data on a particular individual is plentiful, a model can be constructed using that individual's data only, however, as is often the case, the individual data can simply be included with all subjects and split between training and testing. Because across-person sample variability is often greater than within-person variability, model testing with an individual's data in both train and test sets often leads to higher reported accuracy.

However, if the likely application is meant for use in a lab or at-home setting where steps will not be made to collect an individual's data and tailor the model to the individual, such reported results should be discarded. Straightforwardly, if the goal of the developed model is to be applied to people for which it has not been specifically trained, the modelers and reviewers should be emphasizing subject-wise cross-validation. For specific recommendations, subject-wise cross-validation should be primary rather than secondary in an abstract and throughout the narrative of a paper. Subject-wise cross-validation models will necessarily be less accurate than individually-trained models, however, with enough subjects the differences should be diminished. This can be observed in learning curves where the accuracy on the test set is plotted

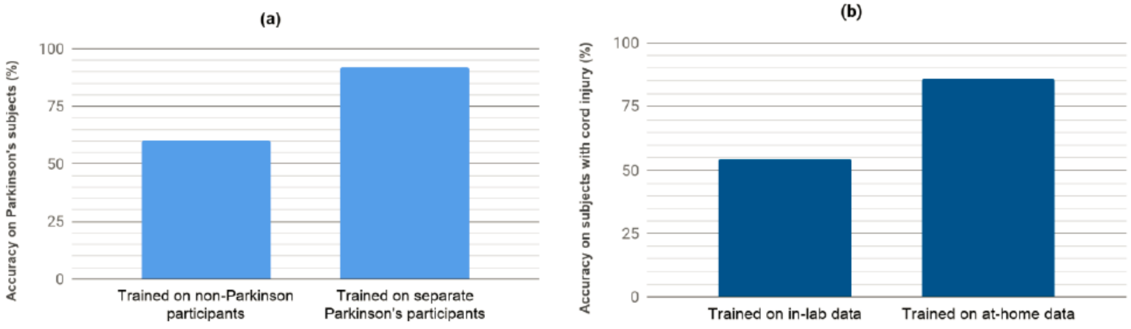

Figure 2: Evidence for benefits of tuning machine learning models to subject populations and context (a) An activity recognition model applied to individuals with Parkinson's disease performs poorly when trained on non-Parkinson's health subjects, but improves by 31.9% when trained on separate Parkinson's subjects [37] and (b) An activity recognition model was testing on data collected at home for incomplete spinal cord injury subject data. By training the model data from other subjects at home, as opposed to similar subject in the instructed lab environment, the accuracy improved by 31.0% [41].

relative to the amount of data - in this case, the number of subjects used. With enough subjects, over time similar individuals may be represented in the data set which would cause the model to converge with an individual or pooled subject model.

5.3 Tailoring to the Population

Trained models may function on subpopulations that differ from the group for which the model was designed. However, in medical application such uses are best validated for the particular population. In the cases in which the model validation demonstrates poor performance, it may be possible to improve that performance by tailoring the model to the specific subpopulation.

For example, when creating an activity recognition system for older individuals or people with movement disorders, unique patterns of movement may be misidentified. A study validating step counters among age groups demonstrated most function well for young adults but tend to undercount among older adults [36]. Population-specific data can be directly used to improve training accuracy. For example, Figure 2(a) shows a model trained on non-Parkinson study participants performed poorly when applied to participants with Parkinson's (60.3% accuracy). However, when the model was trained on separate Parkinson's participants, the model performance dramatically improved (92.2% accuracy) [37]. Similar improvements can be observed in other subject populations including incomplete spinal cord injury subjects [38], lower-limb amputees [39], and even toddlers [40]. A dramatic improvement in accuracy is possible when trained with data from the population to which the model will be applied.

5.4 Tailoring to the Context

Continuing to use activity recognition using wearable devices as an illustration, it is well known that activity recognition systems which are designed using data in the clinic often perform poorly when used outside the clinic. One reason for this is that the types of movements observed when instructed in the clinic differ significantly from the natural movements at home. For example, instructing a person to walk in the clinic leads to a very stereotyped pattern of walking that will be fairly similar across individuals, whereas someone naturally walking at home would have a different gait including more irregular intervals of movement, changes in direction, and irregular cadence. To know how a recognition system would work at home one it is important to also consider at-home collected data for a proper assessment. However, we can do more than simply validate in a particular context; through machine learning we can use data collected in a particular context to increase the accuracy of a model in that context. For example, Figure 2(b) exhibits an activity recognition study involving adult participants with incomplete spinal cord injury, it was observed that accuracy dropped to 54.6% when testing a classifier on at-home activities in a lab setting. But when the classifier was trained on at-home data, the accuracy increased to 85.6% [41]. In general, having data which matches the context in which the data set will be applied enables higher accuracy, particularly in scenarios where data is difficult to acquire.

6 CONCLUSION

It is critical in the practice of medicine that decisions are made accurately. Due to the challenge of thoroughly understanding decisions of complex learning models, proper validation is becoming increasingly critical. We addressed a number of common mistakes ranging from a simple lack of adequate validation to common mistakes like reporting validation error as test error. Poor subject sampling and lack of natural variability in training and testing (variability that is ever-present in medical applications!) are concerns that are simple to state but nuanced to address.

We suggest some methodological improvements directly, such as a greater use of subject-wise cross-validation and nested cross-validation, which should be emphasized in machine learning and AI education. It is also important to be aware of the implicit assumptions in the metrics used, or at least recognize alternate metrics and practical reasons to consider them - e.g. mean error versus log error metrics. Most importantly, in medical contexts the balance between predictive power and the need for natural variability can be simultaneously addressed through proper sampling of populations and context. It is important that models designed for particular groups, particular contexts, or even particular individuals be tested in a way that matches how the developed model will be applied.

In conclusion, we note that by acquiring the data sets with more natural variability tailored to the applied context, avoiding overfitting, and evaluating models by using more appropriate metrics, predictive models in medicine will be more reliable, robust, and replicable. Through this greater accuracy, we hope to see more acceptance of machine learning in healthcare and a corresponding increase in the benefits that quality predictive models can provide for medical practice.

REFERENCES

[1] R. Bhardwaj, A. R. Nambiar, and D. Dutta, “A Study of Machine Learning in Healthcare,” in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Jul. 2017, vol. 2, pp. 236–241.

[2] A. Callahan and N. H. Shah, “Chapter 19 - Machine Learning in Healthcare,” in Key Advances in Clinical Informatics, A. Sheikh, K. M. Cresswell, A. Wright, and D. W. Bates, Eds. Academic Press, 2017, pp. 279–291.

[3] E. Vayena, A. Blasimme, and I. G. Cohen, “Machine learning in medicine: Addressing ethical challenges,” PLoS Med., vol. 15, no. 11, p. e1002689, Nov. 2018.

[4] A. F. Simpao, L. M. Ahumada, J. A. Gálvez, and M. A. Rehman, “A review of analytics and clinical informatics in health care,” J. Med. Syst., vol. 38, no. 4, p. 45, Apr. 2014.

[5] J. H. Chen and S. M. Asch, “Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations,” N. Engl. J. Med., vol. 376, no. 26, pp. 2507–2509, Jun. 2017.

[6] A. Vellido, J. D. Martín-Guerrero, and P. J. G. Lisboa, “Making machine learning models interpretable,” in ESANN, 2012, vol. 12, pp. 163–172.

[7] Z. C. Lipton, “The Mythos of Model Interpretability,” arXiv [cs.LG], Jun. 10, 2016.

[8] P.-H. C. Chen, Y. Liu, and L. Peng, “How to develop machine learning models for healthcare,” Nat. Mater., vol. 18, no. 5, pp. 410–414, May 2019.

[9] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, “Machine learning in cardiovascular medicine: are we there yet?,” Heart, vol. 104, no. 14, pp. 1156–1164, Jul. 2018.

[10] S. Nemati *et al.*, “Monitoring and detecting atrial fibrillation using wearable technology,” in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Aug. 2016, pp. 3394–3397.

[11] A. Bihorac *et al.*, “MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery,” Ann. Surg., vol. 269, no. 4, pp. 652–662, Apr. 2019.

[12] M. Fatima and M. Pasha, “Survey of Machine Learning Algorithms for Disease Diagnostic,” JILSA, vol. 09, no. 01, pp. 1–16, 2017.

[13] Muhammad Aurangzeb Ahmad KenSci Inc. & University of Washington-Tacoma, Seattle, WA, USA, Carly Eckert KenSci Inc. & University of Washington, Seattle, WA, USA, and Ankur Teredesai KenSci Inc. & University of Washington-Tacoma, Seattle, WA, USA, “Interpretable Machine Learning in Healthcare | Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.” https://dl.acm.org/doi/10.1145/3233547.3233667 (accessed Jan. 05, 2020).

[14] B. Goodman and S. Flaxman, “European Union Regulations on Algorithmic Decision Making and a ‘Right to Explanation,”’ AI Magazine, vol. 38, no. 3, pp. 50–57, 2017.

[15] V. Lai and C. Tan, “On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection,” in Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 2019, pp. 29–38.

[16] R. Phillips, K. H. Chang, and S. A. Friedler, “Interpretable Active Learning,” vol. 81, pp. 49–61, 2018.

[17] B. Mittelstadt, C. Russell, and S. Wachter, “Explaining Explanations in AI,” in Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 2019, pp. 279–288.

[18] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, “Prediction Policy Problems,” Am. Econ. Rev., vol. 105, no. 5, pp. 491–495, May 2015.

[19] R. Rao, G. Fung, and R. Rosales, “On the Dangers of Cross-Validation. An Experimental Evaluation,” in Proceedings of the 2008 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2008, pp. 588–596.

[20] C. Zhang, O. Vinyals, R. Munos, and S. Bengio, “A Study on Overfitting in Deep Reinforcement Learning,” arXiv [cs.LG], Apr. 18, 2018.

[21] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, “Reducing Overfitting in Deep Networks by Decorrelating Representations,” arXiv [cs.LG], Nov. 19, 2015.

[22] M. V. Zelkowitz and D. Wallace, “Experimental validation in software engineering,” Information and Software Technology, vol. 39, no. 11, pp. 735–743, Jan. 1997.

[23] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, “The need to approximate the use-case in clinical machine learning,” Gigascience, vol. 6, no. 5, pp. 1–9, May 2017.

[24] B. A. Nosek *et al.*, “SCIENTIFIC STANDARDS. Promoting an open research culture,” Science, vol. 348, no. 6242, pp. 1422–1425, Jun. 2015.

[25] Y. Zhang and Y. Yang, “Cross-validation for selecting a model selection procedure,” J. Econom., vol. 187, no. 1, pp. 95–112, Jul. 2015.

[26] G. Varoquaux, “Cross-validation failure: Small sample sizes lead to large error bars,” Neuroimage, vol. 180, no. Pt A, pp. 68–77, Oct. 2018.

[27] L. Dora, S. Agrawal, R. Panda, and A. Abraham, “Nested cross-validation based adaptive sparse representation algorithm and its application to pathological brain classification,” Expert Syst. Appl., vol. 114, pp. 313–321, Dec. 2018.

[28] D. G. Altman and J. M. Bland, “Statistics Notes: Diagnostic tests 1: sensitivity and specificity,” BMJ, vol. 308, no. 6943. pp. 1552–1552, 1994, doi: 10.1136/bmj.308.6943.1552.

[29] X. Dong, L. Qian, Y. Guan, L. Huang, Q. Yu, and J. Yang, “A multiclass classification method based on deep learning for named entity recognition in electronic medical records,” in 2016 New York Scientific Data Summit (NYSDS), Aug. 2016, pp. 1–10.

[30] Y. Wang, Z. Li, L. Feng, C. Zheng, and W. Zhang, “Automatic Detection of Epilepsy and Seizure Using Multiclass Sparse Extreme Learning Machine Classification,” Comput. Math. Methods Med., vol. 2017, p. 6849360, Jun. 2017.

[31] A. Farooq, S. Anwar, M. Awais, and S. Rehman, “A deep CNN based multi-class classification of Alzheimer’s disease using MRI,” in 2017 IEEE International Conference on Imaging Systems and Techniques (IST), Oct. 2017, pp. 1–6.

[32] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, “Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model,” Sci. Rep., vol. 7, no. 1, p. 4172, Jun. 2017.

[33] E. P. Liski, H. Toutenburg, and G. Trenkler, “Minimum mean square error estimation in linear regression,” J. Stat. Plan. Inference, vol. 37, no. 2, pp. 203–214, Nov. 1993.

[34] G. Brassington, “Mean absolute error and root mean square error: which is the better metric for assessing model performance?,” Apr. 2017, p. 3574.

[35] M. Nicolao, H. Christensen, S. Cunningham, P. Green, and T. Hain, “A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus,” presented at the LREC 2016, Portorož, Slovenia, May 2016, Accessed: Sep. 09, 2020. [Online]. Available: http://eprints.whiterose.ac.uk/109262/.

[36] F. Modave *et al.*, “Mobile Device Accuracy for Step Counting Across Age Groups,” JMIR Mhealth Uhealth, vol. 5, no. 6, p. e88, Jun. 2017.

[37] M. V. Albert, S. Toledo, M. Shapiro, and K. Kording, “Using Mobile Phones for Activity Recognition in Parkinson’s Patients,” Frontiers in Neurology, vol. 3. 2012, doi: 10.3389/fneur.2012.00158.

[38] P. Sok, T. Xiao, Y. Azeze, A. Jayaraman, M. V. Albert, “Activity Recognition for Incomplete Spinal Cord Injury Subjects Using Hidden Markov Models,” IEEE Sensors Journal 18(15), 6369-6374, 2018.

[39] N. Shawen, L. Lonini, C. K. Mummidisetty, I. Shparii, M. V. Albert, K. Kording, A. Jayaraman, “Fall detection in individuals with lower limb amputations using mobile phones: machine learning enhances robustness for real-world applications,” JMIR mHealth uHealth 5(10) e151, 2017.

[40] M. V. Albert, A. Sugianto, K. Nickele, P. Zavos, P. Sindu, M. Ali, S. Kwon, “Hidden Markov model-based activity recognition for toddlers,” Physiological Measurement, 4 (2), March 2020.

[41] M. V. Albert, Y. Azeze, M. Courtois, and A. Jayaraman, “In-lab versus at-home activity recognition in ambulatory subjects with incomplete spinal cord injury,” Journal of NeuroEngineering and Rehabilitation, vol. 14, no. 1. 2017, doi: 10.1186/s12984-017-0222-5.