NTT DATA Services

WHITE PAPER | HEALTHCARE

Assessing Artificial Intelligence for Your Practice

Ensure clinical reliability with increased insights into algorithms

JANUARY 2021

NTT DATA

Table of contents

Executive summary

Advancements in deep learning enabled by the flow of venture capital into artificial intelligence (AI) in imaging analytics have produced algorithms sophisticated enough to be integrated into clinical practice. Early adopters are driving the direction of the market and experiencing the greatest benefit. They're also the first to anticipate and mitigate risk as rapid development and competition in the AI market emerges.

AI will transform diagnostic imaging through productivity gains, improved healthcare screening and reduced diagnostic errors. The aging population demographic and subsequent rise in chronic disease has expanded the need for medical imaging. Specifically, in assisting radiologists, AI could keep costs down while maintaining radiologists' accuracy and speed, creating a much higher standard of care. However, if not validated properly, AI algorithms can increase the risk of consequential systemic errors. This paper explores the validation of algorithms on general and local populations.

A site-specific review of AI algorithms is necessary to determine patient benefit and return on investment and to gauge market competitors effectively. Commonly used metrics such as accuracy, sensitivity and specificity measure performance. Persisting data bias within algorithms can often cause these metrics to become overinflated when deployed in a real-world environment. That's because algorithms developed against a limited source of data will have bias stemming from the small sample size, as well as measurement and systemic prejudice, making any inference on larger populations uncertain. Assessment of these algorithms should be performed on an expansive set of data across different machine types, protocols, patient populations and end users at the site of deployment to reflect the best outcomes.

NTT DATA Services recognizes data at scale as an asset to accelerate the realization of imaging analytics assessment. Our fully managed platform-as-a-service (PaaS) Advocate AI offering provides medical imaging analytics algorithm developers and clients with previously unavailable, curated, anonymized Digital Imaging and Communications in Medicine (DICOM) data sets and an on-demand, purpose-built computational platform at scale for the creation and validation of AI algorithms. NTT DATA Clinical Imaging Insights builds on our Advocate AI tools, services and validated algorithms, and leverages our competencies in integrating clinical systems and spurring clinical workflow adoption. This software-as-a-service (SaaS) offering deploys validated algorithms, developed through Advocate AI or sourced elsewhere, to analyze medical imaging studies at the modality and then integrate the analytic outputs into the radiology workflow for diagnostic reporting in, for example, the institutional picture archive and communication system (PACS) solutions.

Choosing the algorithm that matches the expected performance can be challenging. Algorithms need meticulous validation before application, and each requires consistent vigilance and maintenance throughout its lifecycle. Users of AI tools should understand what type of data set was used to both train and test the algorithm, how it had been piloted with real-word results and if the algorithm is making decisions based on what physicians/AI users consider clinically relevant features. The proper review and transparency of algorithms will lead to the appropriate application, which can provide value to processes of clinical care.

Classification algorithm evaluation techniques

When choosing to adopt AI as part of the clinical workflow, it's important to understand the science behind the development of the AI algorithm and the various metric terms used to describe the AI's performance. At the core, the goal of training and validating AI is to generalize correctly — referring to AI's ability to perform well on unseen data — and yet AI struggles because it "…cannot reason abstractly, does not understand causation and struggles with out-of-distribution generalization."[1] Algorithm evaluation aims to estimate the accuracy of an algorithm on your future (unseen/out-of-sample) data. The transparency of what methods and data sets were used for the AI tools you choose to evaluate requires an understanding of the basics discussed in this paper.

The term accuracy is a common evaluation metric. It's the fraction of predictions the algorithm got correct (number of correct predictions/total number of predictions). Although powerful, this metric may give a false sense of achievement in situations where there's an imbalance in the data used for training. For example, if a cancer detection algorithm is tested on data where 95% of the CT scans represent benign tumors, it would perform at 95% accuracy if it predicted a tumor to be benign 100% of the time.[2] Thus, the algorithm is no better than one that has zero predictive ability to distinguish malignant tumors from benign tumors and can miss important positive cases even with an accuracy that seems acceptable.

Better metrics for predicting potential development bias from data sets used in training are specificity, recall and precision. Specificity, also called the true negative rate, is the proportion of actual negatives identified correctly. More formally, it's the number of all correct negative results divided by the number of all samples that should be labeled negative. Conversely, recall — also called true positive rate or sensitivity — is the proportion of actual positives identified correctly. If an algorithm has a recall of 0.8, it has correctly identified 80% of the malignant tumors in the data set. Precision, also called positive predictive value, describes the proportion of correct positive identifications. More formally, it's the number of correct positive results divided by the number of positive results predicted by the classifier.[2] If a tumor classification algorithm has a precision of 0.8, it's correct 80% of the time when it predicts there is a malignant tumor. Not all vendors share these measures equally, so it's often hard to understand how to compare the alternatives.

There's a balancing act or tension between precision and specificity versus recall and precision. If you increase the number of positively labeled cases of data, you usually decrease the specificity and precision. Quantifying an algorithm's ability to properly identify positive cases is extremely important in healthcare. In diagnostics, false negatives can have dire consequences. In theory, the number of false negatives would be zero if the algorithm always identified a case as positive. However, the AI then doesn't help clinicians differentiate between diagnoses and thus provides no value. This behavior is also true of mitigating false positives.

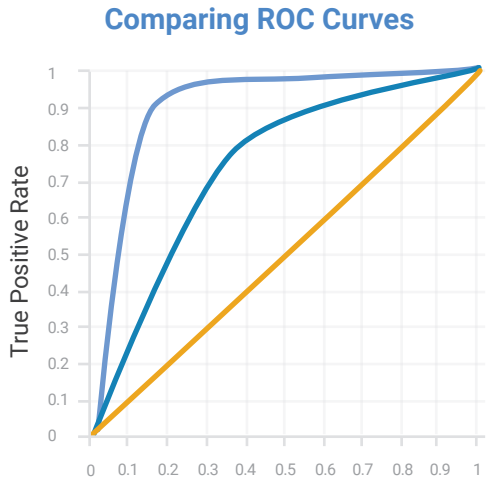

Figure 1. Comparing ROC: The blue line illustrates a strong correlation to the ground truth, and therefore indicates the best AI performance, while the yellow line indicates what is no better than a guess.[3]

Various metrics have been developed to describe this balancing act. In a model of AI performance using receiver operating characteristic (ROC) curve, the true positive rate is plotted along the y-axis and the true negative rate is plotted along the x-axis at different classification thresholds (see Figure 1). Lowering the classification threshold classifies more items as positive, thus increasing both false positives and true positives. This means the shape of the ROC depends on how discriminating the algorithm output is. The area under the ROC curve (AUC) measures the entire two-dimensional area underneath the entire ROC curve.2 AUC ranges in value from 0 to 1. An algorithm whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. Another metric that summarizes over two dimensions is the F1 score, which is the harmonic mean between precision and recall, and it ranges between 0 and 1. The greater the F1 score, the better the performance.

Understanding the impact of data quality

The medical community recognizes a high level of accountability, mitigation of bias and the need for interpretability. In 2019, a consortium of practitioners from across professional healthcare societies produced a statement on the ethics of AI in radiology. A key point within the statement is that "AI in radiology should be appropriately transparent and highly dependable, curtail bias in decision making, and ensure that responsibility and accountability remains with human designers or operators."[4] When making medical decisions, anticipating any unintended consequences or potential bias is the ethical standard.

A predominant theme in the regulatory guidelines for AI is data quality. As expected, bad data is repeatedly cited as the cause for failure of data-dependent initiatives. This is especially true of AI and deep learning, which optimizes its decision ability as it becomes exposed to more data and scenarios. If the data in a training and test isn't representative of the population at large, it will lack generalizability when deployed into production at a single site. Therefore, the performance metrics of an AI algorithm are strongly dependent on the data on which the algorithm is trained. The possible types of bias that can exist within the data are sample bias, measurement bias and prejudicial bias. Even if a data set is extensive, if the algorithm is trained on data that doesn't represent the population or protocols at large, it's destined to fail. External validation of a specific patient population should be considered essential before deployment.

Sample bias occurs if a data set's examples are chosen in a way don't reflect its real-world distribution. A population of data is all the units an algorithm should consider. A recent study shows AI can predict acute kidney failure two days in advance.[5] However, because this algorithm was both trained and tested by data provided by the U.S. Department of Veterans Affairs, the population is overwhelmingly male (approximately 6% of cases are female).[5] If this algorithm was moved to the general population of 50% female, it's unknown how it would perform. Further, if the algorithm performs better on male patients, it could present a difference in the standard of care between male and female patients. These performance uncertainties are impossible to quantify if a better representation of female data is unavailable.

Radiologic AI applications are especially vulnerable to measurement bias, a systematic value distortion that happens when there's an issue with the device used to observe or measure. An example is when the algorithm generalizes distortions of a specific device. An algorithm trained only on data from one manufacturer's system may not generalize to another manufacturer's machines deployed in the same capacity. It's important for the algorithm to be trained and tested on image data from all major device manufacturers. In other words, an algorithm should be device agnostic.

An even more insidious bias that can hinder AI implementation is prejudicial bias, which occurs when the data is influenced by stereotypes coming from within the population. A widely adopted healthcare AI algorithm evaluates which patients will benefit from high-risk care management by assigning patients a risk score. However, a recent article in the Science Journal from the American Association for the Advancement of Science shows the algorithm exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than white patients and improperly characterized.[6] Black populations of patients are being assigned improper risk scores, which could lead to a lower standard of care based on race. The bias exists because the algorithm uses healthcare cost as a proxy for need for care. The algorithm captured the implicit bias that exists in the healthcare system, as well as the correlation between race and income. It's inappropriate for the algorithm to produce results that incorrectly infer and enforce system prejudices. Mitigating prejudicial bias requires insight into the ways that prejudice and stereotyping make their way into data.

So, when comparing algorithms to determine the best option, it's important to not only evaluate the reported precision metrics but also verify that a sufficiently robust data strategy was deployed. Including real-world complexities from a variety of external sources is paramount. The above illustrations of bias are examples of errors that exist in the data itself, separate from the algorithm. Quantitative measurements like AUC, sensitivity and specificity often don't capture these biases because these persist in the test sets. A healthy level of skepticism is important when assessing algorithms. It's imperative to thoroughly understand the training and test data to judge performance metrics. If an algorithm was only tested against a data set where the population was a biased subset, the performance metrics will be inflated.

Algorithm explainability and interpretability

One seminal paper describes a deep neural network that had been trained to diagnose skin cancer and even matched the accuracy of 21 board-certified dermatologists.[7] A year later, the authors noticed that their algorithm was more likely to label an image as malignant cancer if there was a ruler in the image.[8] Dermatologists often only use a ruler to measure the size of a skin lesion if it's particularly concerning. In other words, the data was giving the algorithm hints to the ground truth. If one clinic always includes rulers, the algorithm will over-predict cancer. If one clinic never includes rulers, the algorithm will under-predict cancer. For these reasons, it's important to not only have a diverse data set but also understand the features in the data that are driving the algorithm's decision.

Being aware of the different system biases requires an ability to explain and interpret how AI makes decisions for output. A common criticism of AI solutions is that they're a black box — a metaphor to describe the opaque inner mechanisms of neural networks. The lack of transparency into what's driving the algorithm's behavior is especially

worrying in high-impact fields such as medicine. Clinical rules and best practices require diagnoses and therapeutic decisions to be transparent and clearly explained. The black box metaphor invokes legitimate concerns involving an algorithm's usefulness, reliability, safety and effectiveness in a clinical environment. That's why machine learning "explainability" has become a necessary implementation in the product lifecycle.

Machine learning explainability is the ability to explain in human terms what's going on within the internal mechanics of a machine learning system. Tools such as activation maps, also called saliency maps, can highlight the parts of an image being used by the AI algorithm. There are many other tools, such as LIME and CAM, that visually show the pixels of the image most important for classification. However, it's still an ongoing challenge to develop systems that can explain an AI's reasoning and allow humans to properly interpret AI output. There has yet to be a universal standard to machine learning interpretability. The saliency maps report the features on an image-by-image basis, not a summation of the entire data set. The skin cancer algorithm biased toward images with rulers reported a handful of saliency maps of malignant tumors that highlighted the actual lesion. Explainable AI can give hints to what's of interest and expose AI weaknesses. However, it doesn't completely explain AI in every scenario, nor does it replace good machine learning practices and bias removal.

Explainable AI helps generate insights by allowing the user to probe its behavior on an image-by-image basis. It's important that an AI vendor not only provides an explainability approach but validates it on a real task with user-provided studies. By properly integrating these saliency maps into the workflow, practitioners will gain useful information and context to ensure they're applying the AI algorithm correctly.

AI algorithms have demonstrated lower false positive and false negative read rates when combined with radiologists as opposed to those same metrics when used by an unassisted radiologist.[9] By assisting radiologists, AI offers the capability to address the quality of diagnostic observation while maintaining radiologists' accuracy and speed. In this context, it's important to have the AI decision process that creates focus for radiologists align with the guidelines of the clinical practitioners and the overall practice as well as leverage the descriptive interpretation insights to improve the quality of reporting and, potentially, billing.

Regulatory validation

Regulatory validation is important but not sufficient to determine how AI will perform on your patient data. Even though the U.S. Food and Drug Administration (FDA) has a review process to determine whether AI software is ready for clinical use, you should run your own tests as a best practice.

The majority of AI algorithms the FDA evaluates fall under Class II and are either cleared or not approved through a 510(k) process. For this process, retrospective data is sufficient and clinical trials aren't required (unlike Class III devices). The FDA recommends using a ROC summary performance metric as part of the primary analysis, and sensitivity and specificity as a secondary endpoint to demonstrate effectiveness.[10] However, there are no set thresholds for these performance metrics.

Documentation for these algorithms must be included as part of the submission to confirm adherence to proper machine learning and software practice. In a 2019 white paper[11], the FDA references examples of good machine learning practices as:

- Data acquired in a consistent, clinically relevant and generalizable manner that aligns with the intended use and modification plans
- Relevance of available data to the clinical problem and practice
- Appropriate separation between training, tuning and test data sets
- Appropriate level of clarity of the output and algorithm aimed at users
- Consistently monitored for effectiveness on real-world data

Although these metrics and practices are important tools to use as benchmarks for its algorithms, the FDA's focus is broader, in that its goal is to establish overall safety and effectiveness for the intended use.[11] AI algorithms must prove to not only have high performance but also be clinically meaningful. There must be a valid clinical association between the algorithm's output and the targeted clinical evidence, supported by literature, clinical research or professional society guidelines. This critical relationship isn't captured by analytical techniques such as specificity and sensitivity. Because most of these systems are intended to complement the radiologist, the validation isn't just the performance of the algorithm but how it operates alongside a real-time user. Therefore, the most effective technologies will be those developed and validated with the understanding of the patient and end user.

Regardless of FDA validation, both the Clinical Laboratory Improvement Amendments (CLIA) and the College of American Pathologists (CAP) require that any new test, device or diagnostic undergo validation before being placed into clinical use.[12] The parameters of success are determined by the medical director; neither CAP nor CLIA provides specific guidance on the validation of image-based AI algorithms. However, in 2013, CAP published guidelines on validating whole slide imaging (WSI), and many of those principles can be applied when validating an image analysis algorithm.[12] The following guidelines are particularly relevant when considering validation of any diagnostic digital imaging system:

- ✓ The validation should be appropriate for and applicable to the intended clinical use and the clinical setting of the application in which it will be employed.
- ✓ The validation process should include a sample set that reflects the spectrum and complexity of specimen types and diagnoses likely to be encountered during routine practice.
- ✓ The validation study should closely emulate the real-world clinical environment in which the technology will be used.
- ✓ It's difficult to detect false negatives because they bypass the organization's defenses. A commonly experienced false negative occurs with manufactured synthetic IDs
- ✓ Revalidation is required whenever a significant change is made to any component.

A lot of how technology is evaluated is done so in terms of statistical accuracy and reaching or surpassing human-level performance. That's completely appropriate and necessary to demonstrate efficacy. However, the link to patient outcome and diagnostic accuracy isn't always straightforward. The algorithms must prove to not only have high performance but also be clinically meaningful.

Imaging AI requires vigilance; we can help

AI algorithms, when combined with radiologists, can assist readers like radiologists. Historically, some AI in radiology has shown that when an algorithm isn't calibrated, the tool can actually increase the time to report. Smart adoption of AI will lower costs while maintaining radiologists' accuracy and speed, and create a measurably higher standard of care. Mitigating the risk of AI adoption depends on understanding the performance metrics being reported, the data bias that could mislead an algorithm and AI's overall limitations.

For AI to qualify for adoption, sensitivity and specificity must be competitive with human-level performance. But stellar performance metrics aren't sufficient to forecast whether the algorithm will maintain that performance in outside environments. Just because an algorithm reports 95% specificity on its test data doesn't mean it will hold 95% specificity as the data changes. It's necessary to demand transparency into the population characteristics of the data used in both training and testing. If the population doesn't reflect the site's population and/or real-world deployment is lacking, these performance statistics aren't sufficient to indicate the algorithm's ability to predict future cases. External validation on the clinical user's data is recommended to establish an accurate estimate of generalization. Further, the dynamic nature of data could cause performance degradation even if the algorithm remains constant. Consistent benchmarking is needed to ensure the algorithm functions properly.

AI algorithms depend on, and are improved by, large volumes of high-quality, disease-specific, annotated training data. Curated data sets serve as the basis for testing algorithms for accurate and sophisticated machine-assisted analysis of images, correlation with disease subtypes, and linkage with genetic and metabolic pathways. NTT DATA Advocate AI and Clinical Imaging Insights solutions recognize medical imaging data as an asset that can accelerate the application of machine learning algorithms through rigorous validation against curated, de-identified DICOM data sets at scale. To this end, we created the Nucleus for Unified Clinical Architecture solution based on 20 years of digital medical archiving services, serving more than 1,000 clinical sites and processing more than 20 billion images (over 300 million studies) in the cloud. This anonymized data allows large-scale validation studies to support proper deployment and boost radiologists' confidence in assistive AI.

In addition to providing data as an asset, NTT DATA has the capabilities to address technical challenges when implementing these validation workflows, including:

- The ability to address privacy concerns and regulatory compliance through data governance in the cloud.
- The integration of relevant data from disparate sources and the ability to present results in a meaningful, time- and location-independent way to enable easier tool integration.
- The ability to integrate and anonymize images, DICOM and non-DICOM in a vendor-neutral solution.
- The adoption of multiple solutions without the implementation challenges and administrative overhead associated with vendor-specific integration.

The vendor-neutral image archive, when paired with a vendor-neutral analytics approach, is an asset for increasing top-line revenue, improving quality and productivity, and ensuring better patient care and health. Properly analyzing the state of your health organization's data is a critical step in any adoption strategy. What differentiates NTT DATA is our ability to ensure the value of the data, algorithms and related services to potential clients, as well as to curate and design data, support analytics, create business processes, staff business operations, and develop the metrics to monitor and manage the business.

About the author

Mitchell Goldburgh, Senior Director,
Enterprise Imaging and Analytics, NTT DATA Healthcare

Mitchell is a 35-year veteran of healthcare imaging, holding a variety of roles in the provider segment, as well as healthcare business development positions for public and startup technology companies. He served as co-chair of numerous healthcare standards committees, and authored chapters in academic journals and books on digital imaging. In his current role, Mitchell provides management around NTT DATA's participation in the evolution and adoption of digital imaging in healthcare, driving NTT DATA's resources around the integration of analytics for imaging.

Sources:

1. Rob Toews. "Deep Learning Has Limits. But Its Commercial Impact Has Just Begun." Forbes. February 2020. https://www.forbes.com/sites/robtoews/2020/02/09/deep-learning-has-limits-but-its-commercial-impact-has-just-begun/
2. Google Developers. "Classification | Machine Learning Crash Course." https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall
3. Karen Steward. "Sensitivity vs Specificity." Technology Networks. April 16, 2019. https://www.technologynetworks.com/analysis/articles/sensitivity-vs-specificity-318222
4. J. Raymond Geis, et al. "Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement." Insights into Imaging, vol. 10. October 2019. https://doi.org/10.1148/radiol.2019191586
5. Nenad Tomašev, et al. "A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury." Nature, vol. 572. August 2019. https://www.nature.com/articles/s41586-019-1390-1
6. Ziad Obermeyer, et al. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science, vol. 366. October 2019. https://doi.org/10.1126/science.aax2342
7. Andre Esteva, et al. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." Nature, vol. 542. January 2017. https://www.nature.com/articles/nature21056
8. Akhila Narla, et al. "Automated Classification of Skin Lesions: From Pixels to Practice." Journal of Investigative Dermatology, vol. 138. October 2018. https://doi.org/10.1016/j.jid.2018.06.175
9. Thomas Schaffter, et al. "Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms." JAMA Network Open. March 2020. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2761795
10. Nicholas Peterick. "Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in - Premarket Notification (510(k)) Submissions." U.S. Food & Drug Administration Center for Devices and Radiological Health. January 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology
11. U.S. Food & Drug Administration. "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)." April 2019. https://www.fda.gov/media/122535/download
12. Brent Tan. "How to Validate AI Algorithms in Anatomic Pathology." College of American Pathologists. 2019. https://www.cap.org/member-resources/clinical-informatics-resources/how-to-validate-ai-algorithms-in-anatomic-pathology

Visit **nttdataservices.com** to learn more.

NTT DATA Services, a global digital business and IT services leader, is the largest business unit outside Japan of NTT DATA Corporation and part of NTT Group. With our consultative approach, we leverage deep industry expertise and leading-edge technologies powered by AI, automation and cloud to create practical and scalable solutions that contribute to society and help clients worldwide accelerate their digital journeys.

NTT DATA
Trusted Global Innovator

 0000012021 | 503167-assessing-ai-for-your-practice-white-paper.indd | Rev. 1.0