Radiology: Artificial Intelligence

REVIEW

A Radiology-focused Review of Predictive Uncertainty for AI Interpretability in Computer-assisted Segmentation

Brian McCrindle, BEng • *Katherine Zukotynski, MD, PhD, PEng* • *Thomas E. Doyle, PhD, PEng* • *Michael D. Noseworthy, PhD, PEng*

From the Department of Electrical and Computer Engineering (B.M., T.E.D., M.D.N.), Department of Radiology, Faculty of Health Sciences (K.Z., M.D.N.), and School of Biomedical Engineering (K.Z., T.E.D., M.D.N.), McMaster University, 1280 Main St W, Hamilton, ON, Canada L8S 4L8; and Vector Institute for Artificial Intelligence, Toronto, Canada (T.E.D.). Received January 22, 2021; revision requested March 9; revision received August 9; accepted August 25. **Address correspondence to** M.D.N. (e-mail: *nosewor@mcmaster.ca*).

Supported in part by Defence Research and Development Canada (grant no. CFPMN2–017-McMaster) via the Innovation for Defence Excellence and Security program.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2021; 3(6):e210031 • https://doi.org/10.1148/ryai.2021210031 • Content code: AI

The recent advances and availability of computer hardware, software tools, and massive digital data archives have enabled the rapid development of artificial intelligence (AI) applications. Concerns over whether AI tools can "communicate" decisions to radiologists and primary care physicians is of particular importance because automated clinical decisions can substantially impact patient outcome. A challenge facing the clinical implementation of AI stems from the potential lack of trust clinicians have in these predictive models. This review will expand on the existing literature on interpretability methods for deep learning and review the state-of-the-art methods for predictive uncertainty estimation for computer-assisted segmentation tasks. Last, we discuss how uncertainty can improve predictive performance and model interpretability and can act as a tool to help foster trust.

Artificial intelligence (AI) has seen a resurgence in popularity since the development of deep learning (DL), a method to learn representations within data with multiple levels of abstraction (1). DL frameworks have been widely successful for a variety of applications, including image object recognition and detection tasks where there is a particular interest in applying this technology to interpret complex medical images (2). As modern DL frameworks are structured through multiple hidden layers of network weights, these networks are coined as *black box* models. An important question that arises from such a model is how to trust the prediction of that model, particularly in instances where that prediction can alter clinical outcomes. Even though black box models are not particularly understood, their use in the medical field is analogous to some pharmaceutical drugs, such as thalidomide, that continue to be the standard of care even when clinicians do not fully understand how they work. In contrast, clinicians can estimate the risks associated with a particular drug or procedure, whereas current DL models have yet to provide similar estimates of risk or uncertainty.

Qualitative interpretability methods for DL models have been discussed in the comprehensive review by Reyes et al (3), where there is a strong focus on classification interpretability using saliency maps, local interpretable model-agnostic explanations, and gradient-weighted class activation mapping. Building on this work, we further clarify the semantics of interpretable and explainable models and introduce various methods that estimate predictive uncertainty. Predictive uncertainty is a quantitative value, or set of values, that estimates a model's confidence in its output: A highly confident model would have low uncertainty, and vice versa.

We narrow our focus to the application of predictive uncertainty for segmentation, a particularly important task for which DL has substantial potential to accelerate typical clinical routines. We further emphasize that interpretability is a key factor if the clinical implementation and longevity of such technology is to be successful. Therefore, throughout this review, we discuss why these uncertainty estimates are important for improving model interpretability, where they can fail, and their importance for detecting out-of-distribution (OOD) samples.

Interpretability, Explainability, and Trust

As machine learning (ML) permeates all facets of society, users and stakeholders have sought understandable ML models (4). The two key terms for understandable ML, "interpretability" and "explainability," are used either synonymously or with explicit distinction (3–8). We borrow from Lipton (9) and state that interpretability and explainability are different concepts and should be treated as such. Rather than trying to comprehend the inner workings of a model, as is done with explainable methods, interpretability attempts to understand model predictions at a higher level of abstraction. Current interpretability methods typically act in tandem with a complex model and attempt to communicate the model's decision through either post hoc methods or interactive approaches (8).

The lack of trust in an ML model is typically attributed to a lack of model transparency, which has led to an increase in interpretability research for ML. Some argue that model interpretability is a precursor to trust (10). We emphasize that trust is particularly important in DL, where models are notoriously complex due to the large

Abbreviations

AI = artificial intelligence, DL = deep learning, MC = Monte Carlo, ML = machine learning, NN = neural network, OOD = out of distribution, VI = variational inference

Summary

Interpretable, highly accurate segmentation models have the potential to provide substantial benefit for automated clinical workflows.

Essentials

- Estimating the uncertainty in a model's prediction (predictive uncertainty) can help clinicians quantify, visualize, and communicate model performance.
- Variational inference, Monte Carlo dropout, and ensembles are reliable methods to estimate predictive uncertainty.
- Interpretable artificial intelligence is key for clinical translation of this technology.

Keywords

Segmentation, Quantification, Ethics, Bayesian Network (BN)

number of parameters used to achieve state-of-the-art predictive performance. The cultivation of trust in ML models could benefit from an externalist epistemologic perspective, where trust is rationally justified through proven, repetitive, and reproducible experiences (11), rather than an attempt to understand what is "under the hood."

The 2018 Radiological Society of North America AI Summit emphasized that building trust is a key component for the practical implementation of AI (12). Education and data curation are of top priority to build trust. Model interpretability tools should foster clinicians' trust of computer-based models. Research has shown that individuals are more likely to follow advice when given notions of confidence (13) and that diagnostic accuracy can decrease when radiologists of all levels of expertise are given inaccurate advice (AI or not) (14). Thus, AI should lead to improved clinical outcomes if models can reach higher than expert-level diagnostic performance and indicate predictive confidence.

Interpretability through Uncertainty

Typical DL algorithms do not assign uncertainty estimates with their output predictions. This lack seems counterintuitive, as traditional classification or segmentation (ie, pixel-wise classification) tasks output a probability that a particular object or pixel corresponds to one of the various classes that the network has been trained to identify. This probability often is interpreted erroneously as model confidence, which may lead to confusion when these algorithms are used (15). Consider an extreme case where a sophisticated DL model has randomly initialized weights and has not been trained. In a tumor segmentation task, for example, the sum of the probabilities for a pixel belonging to either of the two possible classes (tumor, not tumor) must sum to 1. Therefore, there can be instances where an untrained network can output high classwise probability where there is no basis to do so. Figure 1 illustrates such a segmentation output. In these cases, predictive uncertainty methods can provide a way to evaluate model confidence at the output, which should be low and independent of the predicted class probability.

For the following discussion, we narrow our focus to provide a foundation for predictive uncertainty, a metric that is less discussed in the existing literature (3).

Predictive Uncertainty

To describe model uncertainty, we provide an illustrative example. Consider a hypothetical model that has been trained on a large dataset of routine brain MR images with the goal of determining the volume's scanning sequence (T1 weighted, T2 weighted, T2 fluid-attenuated inversion-recovery, etc). Presumably, if the model has been trained well, it will correctly distinguish the sequence with high confidence. What would happen if the model were to be exposed to a modality it was not trained with? As Gal et al (18) describe, this question is an example of OOD data. The desired behavior of the model would be to try and provide a reasonable prediction and report the lack of confidence the model has in its output.

The uncertainty associated with a model's prediction can be broken down into three categories: *(a)* sources of random noise within the data (otherwise known as *aleatoric uncertainty*), *(b)* parameter uncertainty (uncertainty in the model's weights), and *(c)* structure uncertainty (what is the best model for the job). The addition of *(b)* and *(c)* is known as the *epistemic uncertainty*, where predictive uncertainty is the addition of the aleatoric and epistemic uncertainties. Epistemic uncertainty can be minimized by training the model with a diverse set of task-relevant data. Alternatively, aleatoric uncertainty has a fundamental limit to which it can be reduced because noise cannot be fully characterized. In situations where the uncertainty is estimated, indications of high or low uncertainty can benefit directly from human intervention while still providing value to the clinician. These uncertainties become particularly useful when visually displayed, as shown in Figure 2.

Uncertainty can be represented in a variety of different ways depending on the application. In a segmentation task, visualizations can quickly communicate areas in which the model has low or high uncertainty and act as a proxy for the quality of the segmentation (19). Uncertainty values can be aggregated to report a single numeric uncertainty estimate to enable clinicians to directly compare predictions. Whether a reported value or set of values is given within a numeric range, as a confidence interval or a standard deviation, the representation of uncertainty should be malleable depending on the task and the individuals interpreting the model's prediction. As such, uncertainty estimates can be standardized in relation to the task or discipline as clinicians see fit.

Estimating predictive uncertainty is indispensable in the case of DL, and there has been noteworthy work to integrate it into prediction pipelines. The methods that produce uncertainty estimates are rooted in Bayesian and frequentist statistics and are described in the following sections.

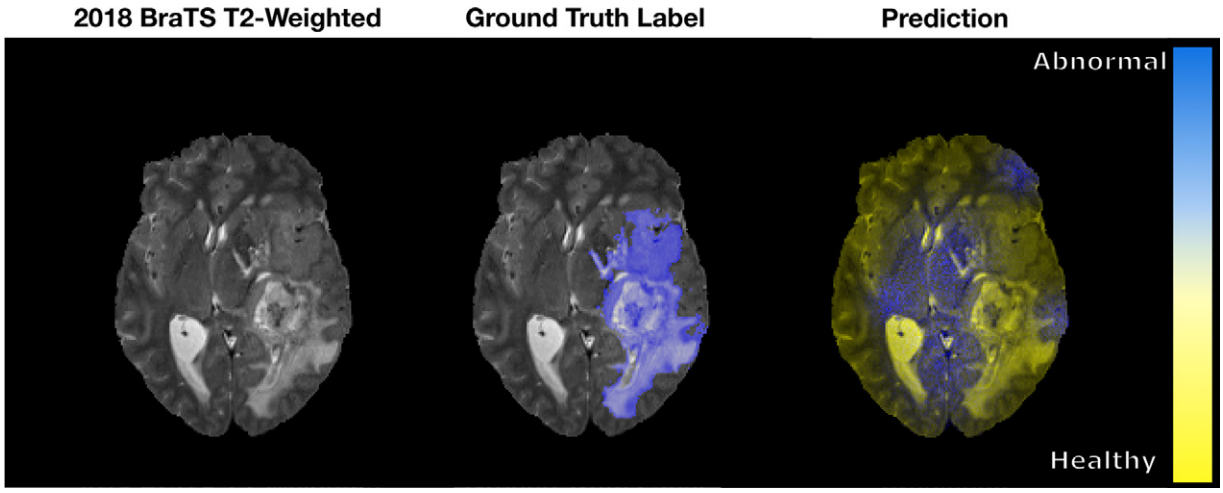

Figure 1: An example of a pixel-wise classification output fused to a sample axial T2-weighted MRI section from the 2018 Medical Image Computing and Computer Assisted Intervention Society Multimodal Brain Tumor Image Segmentation (BraTS) dataset (16,17). The ground truth abnormality is shown in blue. The prediction from the untrained network classifies the abnormal and healthy brain tissues as blue and yellow, respectively. As seen, the network classifies healthy brain tissue as abnormal when there is no basis to do so.

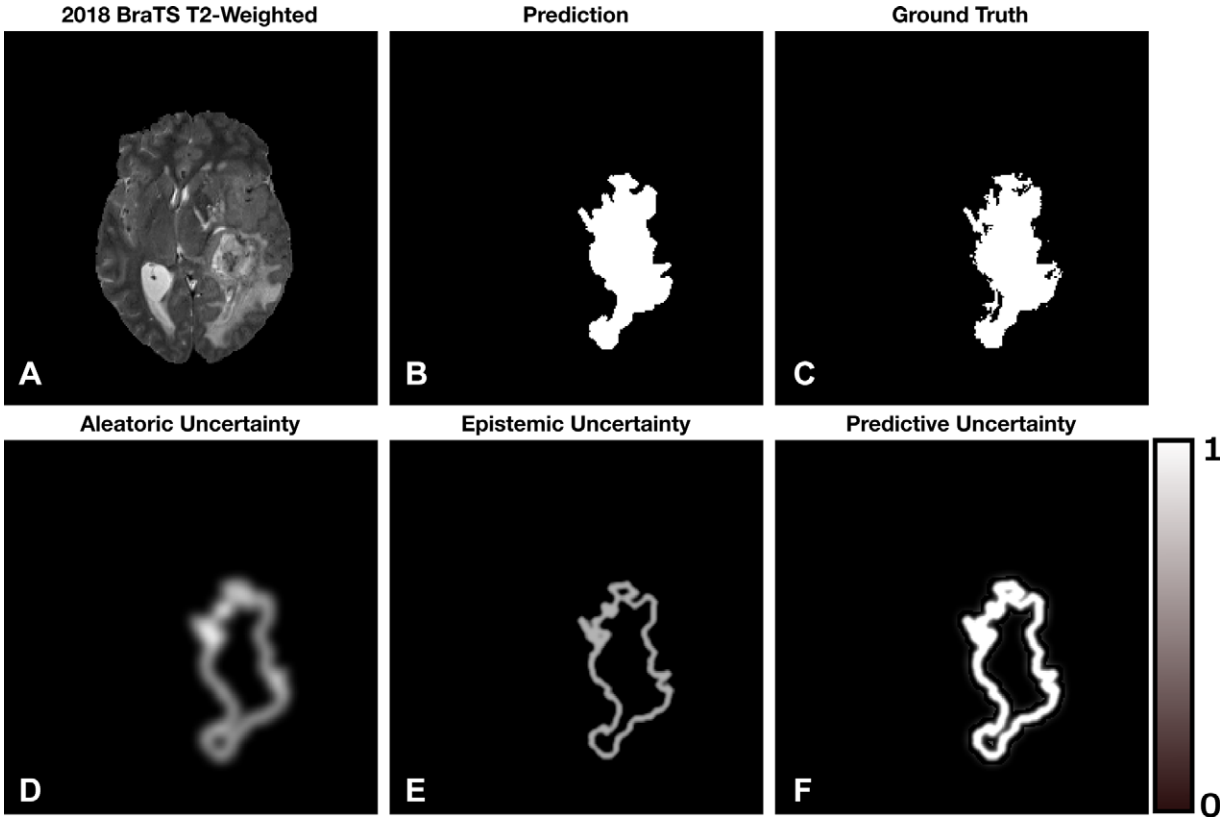

Figure 2: A simulated example of the aleatoric, epistemic, and predictive uncertainties for a segmentation task. Brighter pixels indicate larger uncertainty with range 0–1. **(A)** Axial T2-weighted MRI section from the 2018 Medical Image Computing and Computer Assisted Intervention Society Multimodal Brain Tumor Image Segmentation dataset (16,17). **(B)** The hypothetical segmentation output. **(C)** The ground truth segmentation. **(D)** Aleatoric uncertainty localized to the boundaries of the segmentation. **(E)** Epistemic uncertainty localized to the boundary of the segmentation with less ambiguity compared with the aleatoric uncertainty. **(F)** Predictive uncertainty, which is the addition of (**D**) and (**E**). We notice that the model is confident within the interior of the segmented lesion and less so at the boundary.

Bayesian Neural Networks

Increased data complexity has led to the development of larger neural network (NN) models to obtain state-of-the-art predictive performance. In Bayesian inference, the model parameters are seen as a set of random variables, each possessing an intrinsic probability distribution around its mean. In this formulation, we are ultimately looking to compute a Bayesian model average to determine the probability of the outcome y,

given the data D and network weights ω. This is defined as the predictive probability distribution:

$$p(y|D) = \int p(y|\omega)\, p(\omega|D)\, d\omega$$

This integral becomes intractable with a large number of parameters and is impractical for NNs (18). As such, methods such as variational inference (VI) have been developed to obtain estimates of this integral. To do so, VI postulates an approximate distribution $q(\omega|D)$ that should be distributionally similar to $p(\omega|D)$, the true unknown distribution. To ensure that the approximate distribution is optimal, the difference between $q(\omega|D)$ and $p(\omega|D)$ is measured and iteratively minimized during training (20). Once optimized, the epistemic and aleatoric uncertainties are derived through the variance of the estimated predictive distribution, $p(y|D)$ (21).

A noteworthy limitation of Bayesian modeling is the requirement to inject the necessary prior probability information into the model when determining $q(\omega|D)$. The designer of the model needs to make assumptions about the characteristics of the output distribution where a Gaussian approximation is typically used (22). These assumptions typically hinder the optimization process, which often results in underestimating the predictive uncertainty (20).

VI has been used for regression, classification, and segmentation tasks with varying degrees of success. Kwon et al (21) applied this technique to two multisequence MR image datasets from the 2015 Ischemic Stroke Lesion Segmentation challenge. With VI, Kwon et al were able to build on the initial work completed by Kendall and Gal (23) and proposed a new way to obtain and decompose predictive uncertainty without incorporating additional parameters into the model. The method is able to provide voxel-wise estimations of the aleatoric and epistemic uncertainties, which can then be formed into corresponding visualizations.

VI is a promising Bayesian technique, but it comes with implementation challenges. With the associated complexity of medical image data, optimizing the parameters of the network can be difficult using VI (24).

Monte Carlo Dropout

To obtain results with less computational overhead, methods such as Monte Carlo (MC) dropout have been formulated to provide a simple way to obtain Bayesian-like uncertainty estimates. In the field of DL, dropout refers to a training procedure that randomly suppresses a subset of nodes and their corresponding connections within the network to reduce the chance of coadaptation between layers (25). Figure 3 shows a simple NN with and without dropout. Effectively, dropout forces the network to learn more relevant features, reduces overfitting, and thus improves generalization ability. Initially proposed by Gal and Ghahramani (15), MC dropout uses this concept during both the training and testing procedures of the network. To acquire estimates of the epistemic and aleatoric uncertainties, each input sample is put through N stochastic forward passes. The mean and variance of the set of N predictions is used to determine the uncertainties (18).

Unlike VI, MC dropout requires no prior information to be injected into the model and can obtain an approximation of the output distribution without additional bias. As such, MC dropout can be seen as a frequentist type of solution to estimating predictive uncertainty. With its comparative ease of implementation and Bayesian-like outputs, MC dropout has gained favor within the medical image analysis community compared with its Bayesian counterparts. Nair et al presented the first exploration of multiple uncertainty estimates using MC dropout with a three-dimensional multiple sclerosis lesion segmentation convolutional NN (26). The network was trained with a proprietary, large-scale, multisite, multiscanner multiple sclerosis dataset where voxel-wise and image-wise uncertainty measures were reported. Nair et al showed that by filtering predictions on the basis of its uncertainty, the model's detection accuracy was greatly improved, particularly in the case of small lesions.

Roy et al (27) used MC dropout for full-brain segmentation with voxel-wise uncertainty along with structural uncertainty metrics per chosen anatomic brain structure. Wang et al (28) derived aleatoric uncertainties with MC dropout and test-time augmentation for fetal brain tumor segmentation. In both cases, the uncertainty estimates resulted in improved predictive performance. Although MC dropout provides a simplistic method for obtaining uncertainty, deep ensembles attempt to obtain more accurate predictions by aggregating predictions from many ML models.

Ensemble Methods

The popular method of ensemble-based DL has shown great success for a variety of prediction tasks. By training numerous models, the implementation allows for more robust prediction ability in tandem with OOD stability (29). A DL ensemble aggregates the results from multiple deterministic NNs trained with different parameter initializations. An output probability is determined through the mean of the ensembled predictions, where the variance around the mean is understood as the predictive uncertainty. Each individual prediction before ensembling is seen as a sample from the predictive distribution and as such, the variance between model predictions can be illustrated in a per-voxel fashion.

The generalization ability of an ensemble is often stronger than any of the individual models that compose the framework (30). Therefore, ensembles can result in improved performance over a larger range of relevant inputs and natively include uncertainty measures at the cost of training multiple large networks. Compared with its Bayesian counterparts, the lack of prior assumptions allows for more flexible parameter optimization. Therefore, it is advantageous to use a method such as ensembling to gather more information about the true predictive distribution and obtain greater predictive performance (22).

Ensembling was initially proposed by Lakshminarayanan et al (30) and was evaluated through a series of nonmedical regression and classification benchmarks. Since its inception, the method has been extended for other DL applications. De Fauw et al (31) presented the first clinical application of this method on a large set of three-dimensional optical coherence tomography

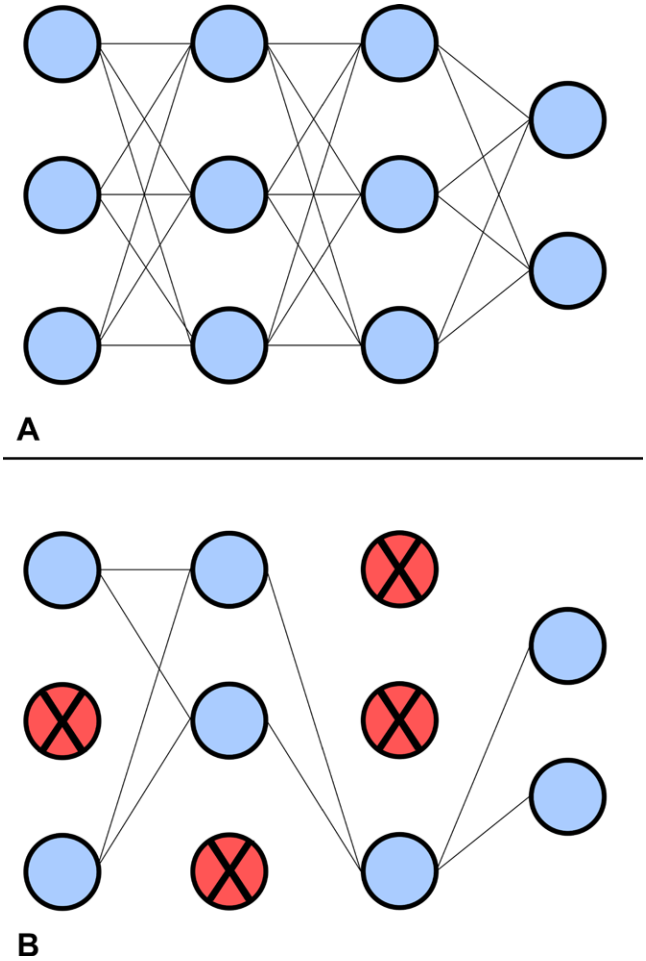

Figure 3: Example of a network **(A)** without and **(B)** with dropout. Nodes and corresponding connections are suppressed during dropout to improve generalization. The probability for a node to be suppressed during dropout is $0 < P < 1$.

scans, showing that ensembling was able to achieve comparable expert-level segmentation performance. Mehrtash et al (32) applied this process to two-dimensional brain, heart, and prostate segmentation tasks for confidence calibration, showing that ensembles performed better in whole-volume and subvolume cases compared with a nonensemble framework. In both studies, the uncertainty estimates were used to improve segmentation in ambiguous regions.

Reliability of Uncertainty Models

When applying uncertainty metrics to a DL model, how can we evaluate which method is optimal? In real-world applications, well-calibrated uncertainty estimates are crucial to determine if a model's output should be trusted. In this case, proper calibration means that the model should output inference probabilities representative of the true likelihood of occurrence (33). More critically, it is important to know which methods work most reliably under dataset shift, a common problem in medical data. Dataset shift means that something has changed between the training, testing, and clinical distributions, where these shifts are normally attributed to changes in population type, acquisition protocols, and/or annotation inconsistencies (34). Therefore, the effective evaluation of predictive uncertainty is most meaningful in the OOD case and, as such, there has been specific work done in an attempt to determine how different methods behave with OOD data.

Ovadia et al (35) presented a large-scale benchmark for a variety of classification problems to investigate the effect of dataset shift on accuracy and uncertainty calibration for VI, MC dropout, and ensemble networks. Trained with nonmedical data, the quality of the uncertainty quickly degraded with increasing dataset shift independent of the model used. Overall, ensembles were consistently seen to perform more accurately across all tested datasets while being the most robust to shifting, even when using a small number of classification models.

Jungo et al (36) suggested that uncertainty estimates should be coupled with an evaluation of model calibration to ensure that the estimates are sensitive to dataset shift. In a recent study evaluating ML accuracy on ImageNet, Shankar et al (37) stated that a model's sensitivity to small, naturally occurring dataset shifts is a performance dimension that is not addressed by current ML benchmarks but is easily handled by humans. To move toward clinical translation, considerable work must be done to ensure that interpretability metrics, such as predictive uncertainty, can capture these shifts.

Discussion

The potential for uncertainty estimates to establish trust instills hope for more informed clinical decisions, but no method comes without limitations. With algorithms reaching expert-level performance, ethical decision-making in the context of diagnostic imaging could become even more difficult when the clinician and algorithm differ in their evaluations and/or recommendations for the patient. The potential for diagnostic disparities between the clinician and the computer will lead to an ethical crossroad: who or which is correct? As Grote and Berens (38) describe, should the clinician trust their initial opinion, default to the algorithm, or bring in additional human supervision? If there is a ground truth that can alleviate this disparity, such as referring to pathologic examination, this should be the default course of action, but in some cases, knowing the ground truth is either impractical or impossible. The action made by the clinician will ultimately rest on the trust and understanding they have in the algorithm, along with their experience within the field.

Unfortunately, the dilemma described has no clear answer in many situations. What should be clear is that predictive models add an extra level of complexity that may or may not improve clinical outcomes, depending on how the clinician interacts with the AI platform. That is, additional information might not always streamline the diagnostic process. In routine clinical cases, it is reasonable to ask whether segmentation models would even need estimates of predictive uncertainty, as the radiologist could indicate if the algorithm matches their own prediction. These situations emphasize the necessity for these models to have improved or equal diagnostic performance compared with radiologists, where poor segmentation models would presumably be avoided regardless of whether uncertainty is provided or not.

Therefore, the efficacy of these predictive models depends greatly on their performance with in-distribution samples, along with the proper calibration and predictive stability in OOD cases. As clinical data distributions shift, uncertainty is indeed a powerful metric to indicate poor model performance. This becomes relevant in automation pipelines where uncertain predictions can be flagged for further review.

Finally, developing causal DL frameworks provides a promising solution for explainable model composition and prediction, but further research must be done to conclude if causality in DL holds up to its proposed claims (34). Furthermore, the norms surrounding AI in health care must be carefully audited, as defaulting to the algorithm can potentially blur the lines of diagnostic accountability (38).

Conclusion

DL has revolutionized the field of ML and has substantial potential to impact clinical workflows. As model interpretability research continues to accelerate, these efforts will slowly increase clinician trust in AI applications. Furthermore, experience with these models will likely play a key role in their clinical translation. Error always exists, be it from a computational model or through expert human interpretation. We must accept some level of risk when implementing intelligent learning machines, but this risk is acceptable only if it is well understood. Collaboration between researchers, clinicians, and medical imaging regulatory bodies will be paramount to set the groundwork for interpretability, trust, and clinical translation in the coming years.

Disclosures of Conflicts of Interest: B.M. Institution received grant from Defence Research and Development Canada (grant number CFPMN2-017-McMaster) for the Innovation for Defence Excellence and Security program presented by the DRDC. **K.Z.** Consultant for the Centre for Probe Development and Commercialization. **T.E.D.** No relevant relationships. **M.D.N.** Data analytics consultant for MCI OneHealth; CEO and cofounder of TBI Finder.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
2. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019;1(6):e271–e297.
3. Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiol Artif Intell 2020;2(3):e190043.
4. Arya V, Bellamy RKE, Chen P-Y, et al. One explanation does not fit all: A Toolkit and taxonomy of AI explainability techniques. arXiv:1909.03012 [preprint] https://arxiv.org/abs/1909.03012. Posted 2019. Accessed November 10, 2019.
5. Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). New York, NY: Association for Computing Machinery, 2020; 648–657.
6. Lakkaraju H, Arsov N, Bastani O. Robust and Stable Black Box Explanations. In: Proceedings of the 37th International Conference on Machine Learning. Vol 119. (Virtual conference): PMLR, 2020; 5628-5638.
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206–215.
8. Selbst AD, Barocas S. The intuitive appeal of explainable machines. Fordham Law Rev 2018;87:1085–1139.
9. Lipton ZC. The mythos of model interpretability. Commun ACM 2018;61(10):36–43.
10. Kim B, Glassman E, Johnson B, Shah J. iBCM : Interactive Bayesian case model empowering humans via intuitive interaction, CSAIL-Technical Rep. (2015) 1–12. https://dspace.mit.edu/handle/1721.1/96315. Accessed August 29, 2020.
11. McLeod C, Zalta E, eds. Trust, Stanford Encycl. Philos. (2020). https://plato.stanford.edu/archives/fall2020/entries/trust/. Accessed September 14, 2020.
12. Chokshi FH, Flanders AE, Prevedello LM, Langlotz CP. Fostering a healthy AI ecosystem for radiology: Conclusions of the 2018 RSNA Summit on AI in Radiology. Radiol Artif Intell 2019;1(2):190021.
13. Gaertig C, Simmons JP. Do people inherently dislike uncertain advice? Psychol Sci 2018;29(4):504–520.
14. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4(1):31.
15. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). 3 (2016) 1651–1660.
16. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993–2024.
17. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 2017;4(1):170117.
18. Gal Y. Uncertainty in Deep Learning [PhD thesis]. Yarin Gal blog. Cambridge Machine Learning Group, 2016. http://mlg.eng.cam.ac.uk/yarin/blog_2248.html. Accessed October 16, 2019.
19. Hoebel K, Andrearczyk V, Beers A, et al. An exploration of uncertainty information for segmentation quality assessment. In: Išgum I, Landman BA, eds. Med. Imaging 2020 Image Process. San Diego, Calif: SPIE, 2020; 381–390.
20. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. J Am Stat Assoc 2017;112(518):859–877.
21. Kwon Y, Won JH, Kim BJ, Paik MC. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Comput Stat Data Anal 2020;142:106816.
22. Wilson AG. Bayesian deep learning and a probabilistic perspective of model construction. International Conference on Machine Learning Tutorial. (2020). https://cims.nyu.edu/~andrewgw/bayesdlicml2020.pdf. Accessed September 14, 2020.
23. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings from the Conference on Advances in Neural Information Processing Systems. Long Beach, Calif: NIPS, 2017; 5575–5585.
24. Shridhar K, Laumann F, Liwicki M. A Comprehensive guide to Bayesian Convolutional Neural Network with Variational Inference. arXiv:1901.02731 [preprint] https://arxiv.org/abs/1901.02731. Posted 2019. Accessed September 22, 2019.
25. Srivastava RSN, Hinton G, Krizhevsky A, Sutskever I. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–1958.
26. Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In: Frangi A, Schnabel J, Davatzikos C, et al, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Lecture Notes in Computer Science, vol 11070. Cham: Springer, 2018; 655–663.
27. Roy AG, Conjeti S, Navab N, Wachinger C. Inherent brain segmentation quality control from fully ConvNet Monte Carlo Sampling. In: Frangi A, Schnabel J, Davatzikos C et al, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Lecture Notes in Computer Science, vol 11070. Cham: Springer, 2018; 664–672.
28. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 2019;335:34–45.
29. Pearce T, Zaki M, Brintrup A, Anastassacos N, Neely A. Uncertainty in Neural Networks: Approximately Bayesian Ensembling. arXiv:1810.05546 [preprint] https://arxiv.org/abs/1810.05546. Posted 2018. Accessed April 2, 2020.
30. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings from the Conference on Advances in Neural Information Processing Systems. Long Beach, Calif: NIPS, 2017.
31. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24(9):1342–1350.
32. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence calibration and predictive uncertainty estimation for deep medical image

segmentation. arXiv:1911.13273 [preprint] http://arxiv.org/abs/1911.13273. Posted 2019. Accessed July 13, 2020.
33. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: 34th Int. Conf. Mach. Learn. ICML 2017, JMLR.org, 2017; 1321–1330.
34. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun 2020;11(1):3673.
35. Ovadia Y, Fertig E, Ren J, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. arXiv:1906.02530 [preprint] http://arxiv.org/abs/1906.02530. Posted 2019. Accessed September 7, 2019.
36. Jungo A, Reyes M. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: Shen D et al, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science. Vol 11765. Cham: Springer, 2019; 48–56.
37. Shankar V, Roelofs R, Mania H, Fang A, Recht B, Schmidt L. Evaluating Machine Accuracy on ImageNet. Int. Conf. Mach. Learn. PMLR 119 (2020).
38. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics 2020;46(3):205–211.