POINT/COUNTERPOINT

Suggestions for topics suitable for these Point/Counterpoint debates should be addressed to Jing Cai, The Hong Kong Polytechnic University, Hong Kong: jing.cai@polyu.edu.hk, and/or Habib Zaidi, Geneva University Hospital, Geneva, Switzerland: habib.zaidi@hcuge.ch, and/or Gerald White, Colorado Associates in Medical Phys, Colorado Springs, CO, United States: gerald.white@mindspring.com. Persons participating in Point/Counterpoint discussions are selected for their knowledge and communicative skill. Their positions for or against a proposition may or may not reflect their personal opinions or the positions of their employers.

Clinical implementation of AI technologies will require interpretable AI models

Xun Jia, Ph.D.
Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75235, USA (Tel: 214-648-5032; E-mail: Xun.Jia@UTSouthwestern.edu)

Lei Ren, Ph.D.
Department of Radiation Oncology, Duke University Medical Center, Durham, NC 27710, USA (Tel: 919-668-0489; E-mail: lei.ren@duke.edu)

Jing Cai, Ph.D., Moderator

(Received 10 October 2019; revised 21 October 2019; accepted for publication 23 October 2019; published 19 November 2019)

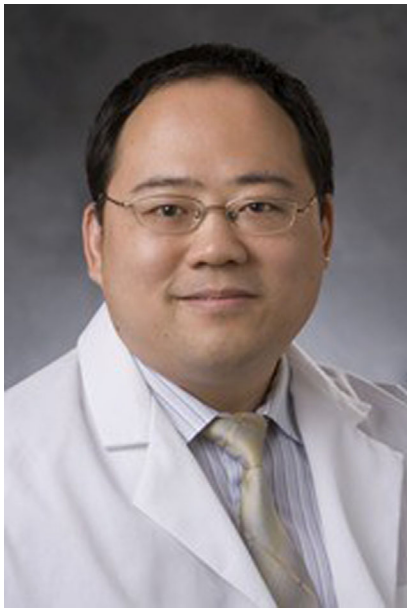
[https://doi.org/10.1002/mp.13891]

OVERVIEW

Artificial intelligence (AI) technologies have been heavily investigated in recent years in various contexts of science and technology. AI models are known to be powerful tools to help us resolve complex problems even though most of us have little understanding of their working principles and even experts may not appreciate which data features drive deep-learning performance in specific applications. These complex models are therefore often referred to as black box models. While some are supportive to the idea that clinical adoption of AI technologies should be only for interpretable models, others believe that it is not required to interpret the AI models as long as they serve the purposes. This is the premise debated in this month's Point/Counterpoint.

Arguing for the Proposition is Xun Jia, Ph.D. Dr. Jia is Associate Professor and Director of Medical Physics research at the Department of Radiation Oncology, University of Texas Southwestern Medical Center (UTSW). He received his Master degree in applied mathematics in 2007 and Ph.D. degree in physics in 2009, both from the University of California Los Angeles. After receiving his postdoctoral training in medical physics from the Department of Radiation Physics and Applied Sciences, University of California San Diego in 2009–2011, he became a faculty in the same department. In 2013, he moved to UTSW. Over the years, Dr. Jia has conducted productive research on low-dose cone beam CT reconstruction, GPU-based Monte Carlo radiation transport simulation, deep-learning based image processing and radiotherapy treatment planning, and development of a preclinical small animal radiation research platform. He has published over 110 peer-reviewed manuscripts. His research has been funded by NIH, the State of Texas, industrial, and charitable funding agencies. Dr. Jia currently serves as an Editorial board member of Physics in Medicine and Biology and an Associate editor of Medical Physics. He is the recipient of John Laughlin Young Scientist Award of American Association of Physicists in Medicine in 2017.

Arguing against the proposition is Lei Ren, Ph.D. Dr. Ren received his Ph.D. degree in Medical Physics from Duke University in 2009, and then worked as a medical physicist in the radiation oncology department at Henry Ford Hospital in Detroit, MI for 2 yr. He joined Duke University as a faculty for both the radiation oncology department and the Medical Physics program in 2011, and is currently an Associate Professor in the department. He is certified by the American Board of Radiology in Therapeutic Medical Physics. In research, Dr. Ren's focus is image-guided radiation therapy (IGRT), including imaging dose reduction, image reconstruction, synthesis, augmentation and registration, 4D imaging, and applications of AI in IGRT. He has published over 50 papers in peer-reviewed journals, including featured articles, and six book chapters. His research has been funded by both NIH and Industry grants and has won multiple awards from AAPM, ASTRO, and ISMRM. Dr. Ren regularly provides



scientific reviews for peer-reviewed journals, conferences, book proposals, and NIH grant applications. He has served in editorial roles for different journals and has been actively involved in multiple committees and annual meeting organization for AAPM and ASTRO. Dr. Ren has mentored seven Ph.D. students, and 16 master students and has received the Mentorship Award from Duke Medical Physics program.

FOR THE PROPOSITION: XUN JIA, PH.D

Opening Statement

A number of novel Artificial Intelligence (AI) studies using deep-neural networks (DNNs) have recently reported impressive performance in various contexts.[1] While implementing these models for routine clinical use is logically the next step, it is probably of equal importance to interpret these models to ensure that the claimed performance and patient safety are sustained.

First of all, data are the foundation of any AI application. An AI model deciphers underlying information buried in the training dataset and interpolates/extrapolates the information to the previously unseen data domain. The sparser the training data points, the riskier it is to interpolate/extrapolate the learned information to new cases. Yet, due to challenges in data collection, it is not uncommon that small-size (~100 or even less) datasets are used in medical physics studies. This is in stark contrast with datasets in many remarkable general image-processing projects, e.g., imageNet with $>10^6$ images.[2] Small datasets, coupled with problems with a high dimensionality, implies that AI problems are often in the so-called small-n (sample size) large-p (problem dimension) regime,[3] where both model training and evaluation are known to have large uncertainties. While it is challenging to construct AI models and comprehensively evaluate them with large datasets, model interpretation is of utmost importance to guide us toward building models with guaranteed performance.

Second, interpretable AI models allow us to identify their application scopes and to ensure that they are applied to the right scopes. Characteristics of the datasets for model construction may have subtle difference from that of the data in subsequent clinical implementations. An AI model may agree well with the training/testing datasets, but lose its performance at the clinical application stage. This is particularly a concern given the large capacity of a DNN model to often fit data nicely including data noise,[4] and the lack of robustness that makes an AI model sensitive to subtle changes in data characteristics.[5]

We also have to be careful about confounding factors that may be hidden in the data and could mislead AI models. For example, a recent study examining DNN-based pneumonia detection on x-ray images revealed that the network was actually trained to identify hospitals with higher prevalence of the disease from input images, and employed this information to make predictions,[6] as opposed to using only anatomical features. Again, interpreting AI models helps avoid mistakes like this, which could be catastrophic, if not detected before clinical implementation.

In summary, while we celebrate the success of AI models, we probably need to understand reasons for this success and to unveil the corresponding model limitations. Interpretability is valuable in overcoming challenges caused by the aforementioned issues to ensure safe and effective applications of AI models, which will ultimately generate positive healthcare impacts.

AGAINST THE PROPOSITION: LEI REN, PH.D

Opening Statement

AI, especially deep learning (DL), has advanced rapidly in the past few years. Several studies have demonstrated that DL outperformed other machine learning techniques with superhuman performance in several clinical tasks.[1] However, DL algorithms are often considered as a black box without clear interpretation, which leads to concerns about prediction bias and robustness against different datasets. Although developing interpretable AI models could address this, I do not think that it should be required for implementing AI technologies for the three main reasons below.

The first reason is that AI can be deployed as a tool, with full human oversight and approval of the end result.[1] In this hybrid AI-human model, physicians verify and correct any errors in the AI prediction. As a result, noninterpretable models can be safely implemented to improve the efficiency and quality of patient care. Take the contouring task in radiotherapy for example. A recent multi-institution study showed that DL improved physicians' tumor contouring accuracy for nasopharyngeal carcinoma and reduced intra and interobserver variations by 36.4% and 54.5% respectively.[7,8] Notably, the physician contouring time was also reduced by nearly 40%. Other examples include AI-assisted diagnosis,[9] automatic treatment planning, etc.[10]

The second reason is that AI can be made safe and effective through proper training and verification methods, instead of developing interpretable models. For example, proper curation of the training, validation, and test datasets to ensure their sufficient size, quality, and independence can reduce bias and enhance robustness of the AI model.[1] Rigorously cross-validating the trained AI model at multi-institutions can verify the

model's repeatability and generalizability. Auditing the black-box algorithms can be used to identify and correct for any prediction bias.[11] Developing lifelong learning models can continuously improve the performance of AI by accommodating new knowledge and data obtained over time.[12]

The final, and perhaps the key reason, is that in many instances, from the patients' perspective, effectiveness is far more important than interpretability. Medicine itself is largely experience-based with many breakthrough treatments that were clearly effective and yet not mechanistically understood initially. For example, in 1847, Dr. Ignaz Semmelweis found that doctors' hand washing with chlorinated lime solutions following autopsies reduced the puerperal fever mortality rates by 90% in obstetric clinics. However, the mechanism of this life-saving practice was not clear until years later when the germ theory was finally confirmed. In general, there is a trade-off between model accuracy and model interpretability.[13] As a patient, would you choose an interpretable mediocre treatment or an unexplained cutting-edge treatment with a better outcome? I believe most patients will opt for the latter since eventually, everything boils down to the result when making clinical decisions.

In summary, lack of interpretability should not be a barrier to implementing effective AI models to improve the quality and efficiency of patient care. Careful data curation, rigorous validation, and human oversight can minimize the risk of less interpretable AI models. As physicians start to gain more experiences in working with AI and understanding its strength and weakness, trust can be built to make AI an indispensable assistant in routine clinical practice.

Rebuttal: Xun Jia, Ph.D

I agree with my opponent that the hybrid human-AI approach helps verify and correct errors. Yet a human cannot completely verify all aspects of a model. Robustness is one example. A recent article in *Science* discussed deep vulnerability of AI models in healthcare: a small change in input can make the model confidently give wrong conclusions.[14] Such phenomena are not few and far between, but exist in models for the diagnosis of diabetic retinopathy from retinal fundoscopy, pneumothorax from chest x-ray, melanoma from dermoscopic photographs,[14] and lung nodule from CT.[15]

While proper training, validation, and test can ensure model performance in theory, this is practically challenging in many medical physics contexts due to limitations in data size and quality. AI models are often built using "free" data collected in routine practice, which are not under well-controlled conditions. Models built using these data risk bias or confounders. Data sharing may overcome certain limitations, but is difficult because of legal and privacy concerns. Model interpretation allows us to incorporate human knowledge to reduce requirements on data.

As for the point that effectiveness is more important for patients than interpretability, I agree, but would emphasize that decision-making based on observed correlations is risky. The doctors' hand washing example brought up by my opponent was a fortunate case in history. Yet we should not forget unfortunate ones. For instance, observational studies discovered lower rates of coronary heart disease (CHD) in women receiving hormone replacement therapy (HRT).[16] However, randomized trials gave the opposite conclusion about 10 yr later.[17] Re-analyses of the observational studies revealed data bias that accounted for the initial incorrect conclusion. Nonetheless, the damage has been made already: for a long time, recommendation to consider HRT was given to women to lower CHD risk. To me, finding reasons behind observed correlations, e.g., the role of germs in the doctors' hand washing example, indeed highlights the need for model interpretation, to confidently enjoy benefits or to prevent harms.

To conclude, several factors may affect observed performance of AI models. Model interpretation is valuable to warrant their clinical benefits and patient safety.

Rebuttal: Lei Ren, Ph.D

My colleague raised several valid points regarding potential limitations of the low-interpretability DL models. Although interpretable models can alleviate these limitations, they often come at a price of sacrificing the prediction accuracy, given the inverse relationship between model interpretability and model accuracy.[13] While interpretation is important and I believe it will eventually come for DL models, our current understanding together with tools, procedures, and safeguards can overcome the limitations of DL and bring immediate real benefits to our patients. This is illustrated in my responses to the three arguments my colleague has made:

1. Lack of sufficient training data to build a robust AI model.

   There have been multiple ways to address the lack of training data. Several techniques were developed to expand the training datasets, including data augmentation, advanced data annotation methods, domain adaptation, and data synthesis.[1] For example, in domain adaptation, transfer learning has been developed to pretrain the DL model using large natural image data before the training using medical data.[18] There have also been ongoing efforts to build large public databases for model training, such as the "ChestX-ray8"[19] and the "DeepLesion"[20] databases provided by the NIH. Distributed learning was also developed to train the models at individual institutions before integrating them into a single model without pooling the data together, which is often challenging in practice.[21]

2. Generalization of the DL model.

   The generalization issue applies to both DL and interpretable models. As stated in the opening statement, the generalization of the DL model can be improved through comprehensive and rigorous training and cross-validation of the model using multi-institution datasets. Also, adversarial attacks can be used to test the robustness of the DL

model against data variations, and defense strategies can be designed accordingly to address any weakness.[22]

3. Potential bias in the DL prediction.

As mentioned above, the bias in the DL prediction can be detected by auditing the results to analyze the effects of various factors on the prediction.[11] In fact, the example my colleague mentioned demonstrated exactly how auditing was able to catch bias in the DL prediction so that it can be addressed before implementation.

Besides the technical tools mentioned above, human oversight and approval of the end result provide additional safeguards to ensure safe and effective implementations of DL models.

In conclusion, with a superior prediction accuracy than interpretable models, DL can be effectively implemented with proper tools, procedures, and safeguards to aid the clinicians in achieving higher quality and better efficiency in patient care.

CONFLICTS OF INTEREST

Dr. Jia and Dr. Ren have no relevant conflict of interest.

REFERENCES

1. Sahiner B, Pezeshk A, Hadjiiski LM, et al. Deep learning in medical imaging and radiation therapy. *Med Phys*. 2019;46:e1–e36.
2. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211–252.
3. Bishop CM. *Pattern Recognition and Machine Learning*. Berlin: Springer (India) Private Limited; 2013.
4. Zhang C, Bengio S, Hardt M, Recht B,Vinyals OJ. Understanding deep learning requires rethinking generalization; 2016.
5. Goodfellow IJ, Shlens J, Szegedy C.Explaining and harnessing adversarial examples. arXiv preprint arXiv:14126572; 2014.
6. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine Med*. 2018;15:e1002683.
7. Lin L, Dou Q, Jin YM, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology*. 2019;291:677–686.
8. Chang Z, Will AI. Improve tumor delineation accuracy for radiation therapy?. *Radiology*. 2019;291:687–688.
9. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350.
10. Nguyen D, Long T, Jia X, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep*. 2019;9:1076.
11. Adler P, Falk C, Friedler SA, et al. Auditing black-box models for indirect influence. *Knowl Inf Syst*. 2018;54:95–122.
12. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual life long learning with neural networks: A review. *Neural Netw*. 2019;113:54–71.
13. Ahmad MA, Eckert C, Teredesai A, McKelvey G. Featured Article: Interpretable Machine Learning in Healthcare. IEEE Intelligent Informatics Bulletin.2018;19(1).
14. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. 2019;363:1287–1289.
15. Tsai M-Y, Shen C, Jia X.Evaluating Robustness of Deep Learning Based Lung Nodule Classification. AAPM annual meeting.2019;Science council session:TU-HI-SAN2-11.
16. Grady D, Rubin SM, Petitti DB, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med*. 1992;117:1016–1037.
17. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA*. 1998;280:605–613.
18. Deng J, Dong W, Socher R, Li L-J, Li K,Fei-Fei L.Imagenet: a largescale hierarchical image database. IEEE conference on computer vision and pattern recognition; 2009:248-255.
19. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM.ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017:2097-2106.
20. Yan K, Wang X, Lu L, Summers RM. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J Med Imaging (Bellingham)*. 2018;5:036501.
21. Jochems A, Deist TM, van Soest J, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - a real life proof of concept. *Radiother Oncol*. 2016;121:459–467.
22. Yuan X, He P, Zhu Q, Li X. Adversarial examples: attacks and defenses for deep learning. *IEEE Trans Neural Netw Learn Syst*. 2019;30:2805–2824.