Radiology: Artificial Intelligence

COMMENTARY

The Quest for Generalizability in Radiomics

Philip O. Alderson, MD

Philip O. Alderson, MD, is dean emeritus of the School of Medicine, Saint Louis University, St Louis, Mo. He served as dean from 2008 to 2016. He previously was a faculty member at Columbia University Medical Center in New York, NY (1980–2008) where he was James Picker Professor and chair of the department of radiology (1988–2008). He also served on the faculty at the Johns Hopkins Medical Institutions, Baltimore, Md (1976–1980). He is a past president of the ABR, ARRS, AUR, APDR, the Academy of Radiology Research, and the Fleischner Society, and is a fellow of the ACR, AAAS, and AIMBE.

In recent years, there have been numerous efforts to quantify subvisual structural features present in medical images generated by CT, MRI, PET, and other types of digital imaging. The hope is that quantitative information obtained from subvisual features will extend diagnostic utility beyond the capabilities of traditional visual inspection. Using terminology similar to that used in genomics and other "-omics," this group of approaches is termed *radiomics*, and the derived features themselves are known as *radiomic features* (RFs).

There are many types of RFs. RF extraction methods include categories of statistical, filtering, and morphologic features. Many statistical representations of pixel intensity changes across objects of interest exist and are said to be measures of texture. The Image Biomarker Standardization Initiative has defined many such features to help develop reproducible classifiers (1). Additional feature variations have been described through the application of mathematical filters, such as the wavelet distribution or fractal dimensions, or various mathematical transforms.

All told, there are hundreds of computational features that can be applied to analyze the ultrastructure of imaged objects. There is no way a priori to know which feature or features might work best to differentiate actionable components associated with a particular disorder or clinical question, so methods have been developed to help make this selection. This process of identifying the best RFs to apply to specific problems is complicated by the fact that RFs that might be consistent biomarkers may be rendered inconsistent by differences in primary signal quality, signal-to-noise ratios, resolution, artifacts, and other machine attributes in different scanners or in the same scanner over time. This may cause the RFs derived from these instruments to vary, and this can lead to inaccurate portrayal of important biologic features. Accordingly, it has been difficult to use radiomics in the conduct of multi-institutional studies or even in the conduct of longitudinal studies in the same facility. Efforts to control this underlying technical variability include the application of a series of strict quality control measures (2), including phantom studies (3). Strict quality control potentially limits the contributions of underlying technical variations. Quantitative artifacts in image acquisition and processing documented by imaging phantoms can be used to adjust for machine-related discrepancies.

Some investigators have used machine learning approaches to minimize manufacturer-based RF variabilities. The article by Marcadent et al in this issue of *Radiology: Artificial Intelligence* proposes a way to harmonize RFs in digital chest radiographs derived from disparate manufacturers using generative adversarial networks (GANs) (4). GANs are a relatively new convolutional neural network (CNN) architecture; there are numerous GAN models. Marcadent et al propose a cycle GAN–based approach to image translation between different radiographic acquisition systems (4). Cycle GANs work in the absence of a detailed paired image to achieve image-to-image translation (5). In this case, a cycle GAN has two linked CNNs, a generator and a discriminator, where the goal of the generator is to create more realistic "fakes" and the role of the discriminator is to get better at being able to identify the fake from real.

Marcadent et al (4) used the two linked CNNs to operate in coordinated cycles such that the feature sets of the original image from the CNN generator are learned by the discriminator and the data of both the generator and discriminator CNN can be blended. This type of feature blending can be easily appreciated by visual inspection of cycle GAN–produced photographs of different animals seeming to become a new blended animal (eg, zebras to horses) (5). Marcadent et al applied this approach to the digital chest radiographs acquired from two different manufacturers (4). In this approach, "fake" images are generated using images from manufacturer A ("source" domain) that mimic the texture of images from manufacturer B ("target" domain). They showed that they could create datasets that were no longer distinguishable digitally and that yielded digital radiographs that were not distinguishable by radiologists even though those radiologists were able to tell the manufacturers apart before the cycle GAN operation. They also showed that after the cycle GAN operation

From the Office of Student Affairs, Saint Louis University, 1402 S Grand Blvd, Caroline 100, St Louis, MO 63104. Received April 18, 2020; revision requested April 23; revision received April 24; accepted May 13. **Address correspondence to** the author (e-mail: *philip.alderson@health.slu.edu*).

See also the article by Marcadent et al in this issue. Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(3):e200068 • https://doi.org/10.1148/ryai.2020200068 • Content codes: IN CA CH • ©RSNA, 2020

their method retained the ability to properly classify abnormal findings on the original chest radiographs. In their study, the abnormality was the presence of congestive heart failure. Thus, they concluded that the application of cycle GANs might be a method to significantly reduce the impact of variabilities in machine signal as a source of inaccuracies in multimanufacturer, multisite, and sequential radiomics.

Others also have addressed the issue of radiomic generalizability. Dercle et al used machine learning–based feature reduction and ranking along with strong quality control to discover reproducible RFs derived from abdominal CT images that were acquired in multiple centers in patients who had colorectal cancer with liver metastases (6). Their chosen RFs were able to be applied to accurately classify patients who would have longer overall survival and which patients were more likely to respond to certain types of chemotherapy. Ohrlac et al used a postreconstruction Bayesian maximum likelihood transformation process (COMBAT) to harmonize the RFs of independently acquired PET/CT data and verified the success of this method in a study of data acquired from CT phantoms (7,8). Choe et al used deep learning–based image kernel conversion to improve the reproducibility of 21% of more than 700 RFs that they were using in chest CT studies of pulmonary nodules and masses (9). These CT scans were acquired over a period of months using the same CT scanner.

Whether RFs are rendered generalizable by using machine learning to select the most valuable variables, by Bayesian transformation, kernel conversion, cycle GANS, or some other method, in a semantic sense all might be described as methods of RF “harmonization.” Time will tell which of these methods or other new approaches prove the most sustainable and useful. Another possibility is that differing RF harmonization approaches will be found to be most successful in differing types of imaging data to which their approach is particularly well suited. Whether the expertise to apply any of these approaches will be widespread or will be available only in select specialized radiomics processing centers remains to be seen.

The cycle GANs approach advocated by Marcadent et al (4) needs independent verification and that verification needs to include studies in CT, which is the modality in which radiomics measurement currently is most frequently applied. Their study suggests that cycle GANs might allow radiomics from independently acquired chest radiographs to have clinical value. Whether or not that proves to be true, chest radiography in the current study has served as a suitable model in which to explore the utility of blending cycle GAN features to promote radiomics generalizability.

Disclosures of Conflicts of Interest: P.O.A. disclosed no relevant relationships.

References

1. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295(2):328–338.
2. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14(12):749–762.
3. Mackin D, Fave X, Zhang L, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 2015;50(11):757–765.
4. Marcadent S, Hofmeister J, Preti MJ, Martin SP, Van De Ville D, Montet X. Generative adversarial networks improve the reproducibility and discriminate power of radiomics features. Radiol Artif Intell 2020;2(3):e190035.
5. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. ArXiv:1703.10593 version 6 [cs.CV] [preprint] https://arxiv.org/abs/1703.10593. Posted March 30, 2017. Accessed April 2020.
6. Dercle L, Lu L, Schwartz LH, et al. Radiomics response signature for identification of metastatic colorectal cancer sensitive to therapies targeting EGFR pathway. J Natl Cancer Inst 2020 Feb 4 [Epub ahead of print].
7. Orlhac F, Boughdad S, Philippe C, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med 2018;59(8):1321–1328.
8. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology 2019;291(1):53–59.
9. Choe J, Lee SM, Do KH, et al. Deep learning-based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. Radiology 2019;292(2):365–373.