Kidney Fibrosis Assessment by CT Using Machine Learning

Kevin V. Lemley

An increasing number of applications of machine learning to nephrology are on the basis of convolutional neural networks (CNNs), a deep learning neural net architecture often applied to visual data. Many such tools are being applied to digitized biopsy whole-slide images for purposes of segmentation of those images into pathologic primitives, such as glomeruli, blood vessels, or tubules. The potential uses of CNN in machine learning are, however, much broader. In fact, an image-based CNN may be trained on any annotated data, including datasets with functional rather than structural annotation (1), in a process referred to as supervised learning. The annotation provides the “ground truth” against which the CNN is trained and optimized and ultimately, against which it is tested.

Chantaduly et al. (2) present in this issue of K360 the results of a retrospective pilot study (Artificial Intelligence in Renal Scarring) using machine learning to classify computerized tomographic (CT) images of native and transplanted kidneys as showing either severe ($\geq$50%) or mild to moderate (<50%) interstitial fibrosis and tubular atrophy (IFTA). In this case, the ground truth was obtained from biopsy reports of kidney biopsies performed within 6 months of the imaging studies. The justification for developing a noninvasive tool to assess IFTA is the desire to avoid biopsy complications, particularly when the presence of advanced degrees of IFTA means that the biopsy results would have little influence on prognosis or treatment. Interstitial fibrosis has long been appreciated to predict long-term kidney function outcomes (3).

The study was performed on 152 CT scans from 92 patients. About half the patients had a kidney transplant. Seventy-seven of 92 patients had CKD as defined by an eGFR $<$60 ml/min per 1.73 m$^2$. These scans included images of 300 native kidneys and 76 transplanted kidneys; 79% of scans did not use contrast. The native kidneys scanned in patients with transplants were automatically scored as severe IFTA, as they were not biopsied. A sample of biopsies was reviewed by a pathologist to make sure that the biopsy reports (the source of ground truth) accurately represented the biopsy findings. Discordant cases were rescoring by another pathologist. There was a reasonable spread of mild, moderate, and severe IFTA in the biopsies.

Two different CNN architectures were used on the CT images after manual whole-kidney segmentation. A global slice by slice classifier categorized each two-dimensional section’s image as showing either mild to moderate or severe IFTA (or nonkidney). The other classifier made the same distinction at a voxel level. The final kidney-level classification was made by “majority rule” of the kidney slices or pixels: that is, if the majority of slices were classified as severe, that would be the overall classification of that kidney. Five-fold cross validation with an 80%/20% split was used for model validation. No held-out independent test set of images was used for final testing. Instead, models trained on the Philips and GE scanners were evaluated on images from the Siemens scanner. Performance of the CNNs was assessed by accuracy, sensitivity, specificity, and positive and negative predictive values. In addition, the area under the receiver operating characteristic curves (AUC) was also calculated. Several subcohorts were analyzed as a way to assess for possible confounding features (e.g., use of contrast or not in the CT).

The two CNN approaches gave generally comparable and quite good performance (AUCs about 0.92). On the other hand, the voxel-based approach, for example, missed severe IFTA in about 15% of cases. What constitutes an acceptable level of accuracy for a diagnostic tool depends on several factors, such as the relative consequences of missing a positive condition or of falsely diagnosing the presence of one. If the practical application of this tool would be a decision to defer biopsy in the case of severe IFTA, this means that 15% of patients would be subjected to a biopsy when it was not indicated. Depending on the discriminatory performance of alternative predictive schemes (such as clinical features, other imaging indicators such as ultrasound-determined kidney size, etc.), this may be considered good performance.

Despite relatively good predictive performance, the described CT/CNN method is not meant to be used in a vacuum. Although avoiding a needless biopsy may be a commendable goal, there will certainly be occasions where—despite what appears to be severe IFTA—the chance for greater certainty with respect to
an unclear diagnosis may justify a kidney biopsy, a situation the authors acknowledge.

CT angiography is often used in the predonation evaluation of vasculature and the urinary collecting system of potential kidney donors (4). One reason to try to refine the CNN to distinguish lower levels of IFTA would be to have the opportunity to take a second look at apparently suitable donors who seem to have higher than expected levels of IFTA on their screening CT. Levels of concern in this setting may certainly be IFTA levels <50% or even <25%. It should be feasible to study this with the addition of more cases, specifically more cases with an eGFR >60 ml/min per 1.73 m², which would also allow a held-out final test set of images to be developed. Inclusion of auxiliary clinical or laboratory parameters, such as eGFR, may improve the performance of the CT-based CNN.

Disclosures
K.V. Lemley reports honoraria from the National Institutes of Health for study section service and expert panel service and scientific advisor or membership with the National Institute of Diabetes and Digestive and Kidney Diseases and the editorial boards for *JASN, Kidney International*, and *Pediatric Nephrology*.

Funding
None.

Acknowledgments
The content of this article reflects the personal experience and views of the author(s) and should not be considered medical advice or recommendation. The content does not reflect the views or opinions of the American Society of Nephrology (ASN) or Kidney360. Responsibility for the information and views expressed herein lies entirely with the author(s).

**Author Contributions**
K.V. Lemley conceptualized the study and wrote the original draft.

**References**
1. Kolachalama VB, Singh P, Lin CQ, Mun D, Belghasem ME, Henderson JM, Francis JM, Salant DJ, Chitalia VC: Association of pathological fibrosis with renal survival using deep neural networks. *Kidney Int Rep* 3: 464–475, 2018 https://doi.org/10.1016/j.ekir.2017.11.002
2. Chantaduly C, Troutt HR, Perez Reyes KA, Zuckerman JE, Chang PD, Lau WL: Artificial intelligence assessment of renal scarring (AIRS Study). *Kidney360* 3: 83–90, 2022 https://doi.org/10.34067/KID.0003662021
3. Bohle A, Mackensen-Haen S, von Gise H, Grund KE, Wehrmann M, Batz C, Bogenschütz O, Schmitt H, Nagy J, Müller C, Müller G: The consequences of tubulo-interstitial changes for renal function in glomerulopathies. A morphometric and cytological analysis. *Pathol Res Pract* 186: 135–144, 1990 https://doi.org/10.1016/S0344-0338(11)81021-6
4. Vernuccio F, Gondalia R, Churchill S, Bashir MR, Marin D: CT evaluation of the renal donor and recipient. *Abdom Radiol* (NY) 43: 2574–2588, 2018 https://doi.org/10.1007/s00261-018-1508-1

**Received:** November 29, 2021 **Accepted:** December 2, 2021

See related article, “Artificial Intelligence Assessment of Renal Scarring (AIRS Study),” on pages 83–90.