

Stratified Iterative Learning for Kidney Cortex and Medulla Segmentation Across Healthy and Pathological Kidneys on Contrast-Enhanced CT

Jianfei Liu, Tejas Sudharshan Mathai, and Ronald M. Summers

Radiology and Imaging Sciences, National Institutes of Health Clinical Center,
Bethesda, MD 20892, USA
jianfei.liu@nih.gov

Abstract. Accurate kidney cortex and medulla segmentation enables compartment-specific assessment of renal pathology that cannot be captured by whole-kidney analysis, including cortical changes in chronic kidney disease, corticomedullary alterations in diabetic kidney disease, and compartment-specific fibrosis. However, most existing medical image segmentation methods focus on whole-kidney segmentation, with relatively few studies addressing cortex and medulla segmentation. Furthermore, these studies rarely consider pathological kidneys containing lesions. In this work, we present a stratified iterative learning framework for kidney cortex and medulla segmentation on contrast-enhanced CT scans. The framework progressively incorporates increasingly challenging cases, beginning with healthy kidneys and subsequently adding pathological kidneys with lesions. By leveraging model predictions from simpler cases to facilitate the annotation and refinement of more complex cases, the proposed strategy provides a practical approach for constructing training data across progressively challenging pathological cases. Validation on two external datasets demonstrated robust performance, achieving Dice similarity coefficients exceeding 97.2% for cortex segmentation and 94.7% for medulla segmentation, even in kidneys containing lesions. These results demonstrate the potential of the proposed framework for compartment-specific analysis of renal diseases.

Keywords: Iterative learning · Kidney Cortex and medulla · Tumor · Cysts · nnU-Net

1 Introduction

The kidney cortex and medulla have distinct physiological functions and are differentially affected by renal disease, motivating compartment-specific assessment of kidney health [3]. Quantitative measurements of these compartments, such as cortical thickness and compartmental volume, may provide valuable imaging biomarkers for renal dysfunction. For example, cortical thinning is associated with declining renal function and greater chronic kidney disease severity, highlighting the importance of cortex-specific analysis [1]. In diabetic kidney

disease, structural changes can occur in both the cortex and medulla, making compartment-specific assessment valuable for disease characterization [13]. Similarly, renal fibrosis may affect kidney compartments differently, highlighting the need for compartment-specific analysis [5].

Despite the importance of compartment-specific assessment, existing research has largely focused on whole-kidney segmentation [2, 11], while relatively few studies have investigated segmentation of the kidney cortex and medulla. One of the earliest approaches, CorteXpert, employed a model-based framework with cortex model adaptation and graph-search optimization for automatic renal cortex segmentation from contrast-enhanced CT scans [12]. More recently, deep learning methods based on 3D U-Net and nnU-Net [6] have been developed for automated segmentation of the kidney cortex and medulla and have been applied to quantitative assessment of renal function in chronic kidney disease [8, 10]. However, these supervised approaches require extensive manual annotation and have primarily been evaluated on healthy kidneys, with limited validation on pathological kidneys containing lesions that are frequently encountered.

This work presents a stratified iterative learning framework that progressively leverages models trained on simpler cases to facilitate cortex and medulla segmentation in challenging kidneys with lesions while maintaining high segmentation accuracy. An initial nnU-Net model [6] was trained using labeled cases without lesions. More challenging cases were then progressively incorporated into the training set by applying the current model to generate preliminary segmentations, followed by targeted manual refinement. The segmentation model was iteratively updated through four learning iterations. The final model was validated on two external datasets and demonstrated high segmentation accuracy for both healthy and pathological kidneys.

2 Method

2.1 Data Collection

A total of 176 contrast-enhanced CT scans from five datasets were collected for training and evaluation of kidney cortex and medulla segmentation models. All scans were acquired during either the arterial or portal venous contrast-enhanced phase. The Sarcopenia, Prostate, and Renal Hypertension datasets were collected at the National Institutes of Health Clinical Center (NIHCC) as part of research studies on sarcopenia, prostate cancer, and renal hypertension, respectively. A subset of the KiTS23 dataset, originally acquired at the University of Minnesota (UM) for the 2023 Kidney and Kidney Tumor Segmentation Challenge (KiTS23), was included for training. KiTS23 consists of contrast-enhanced abdominal CT scans with expert annotations of kidneys, tumors, and cysts [4]. A subset of the SAROS dataset, obtained from The Cancer Imaging Archive (TCIA), was included for independent evaluation [7]. Together, these datasets encompass a broad age range and a relatively balanced sex distribution (Table 1). The Sarcopenia, Prostate, and KiTS23 datasets formed the training cohort, with 72 scans containing kidney lesions and 74 without kidney

Table 1: Patient demographics of the five datasets included in this study (age unavailable for SAROS).

Dataset	Training			Testing	
	Sarcopenia	Prostate	KITS23	Renal Hypertension	SAROS
Scans (n)	84	12	50	10	20
Age (years)	66.6 ± 6.1 (59-87)	67.4 ± 9.4 (56-85)	58.5 ± 12.7 (20-83)	64.8 ± 8.5 (49-77)	-
Male (n)	45	12	27	8	0
Female (n)	39	0	23	2	20
Data Source	NIHCC	NIHCC	UM	NIHCC	TCIA

lesions. The Renal Hypertension and SAROS datasets were used for testing. Among the Renal Hypertension scans, 3 contained kidney lesions, while 7 of the 20 SAROS scans contained kidney lesions.

In addition to labeling the kidney cortex and medulla in both kidneys, several auxiliary anatomical structures were included to facilitate segmentation. The segmentation task comprised 13 classes: (1) body, (2) visceral fat, (3) subcutaneous fat, (4) skeletal muscle, (5) left kidney cortex, (6) right kidney cortex, (7) left kidney medulla, (8) right kidney medulla, (9) liver, (10) spleen, (11) left autochthonous muscle, (12) right autochthonous muscle, and (13) kidney lesion. The body class included all remaining body regions not assigned to any of the other 12 classes. Thus, kidney cortex and medulla segmentation was formulated as part of a 13-class segmentation task. TotalSegmentator [11] was used to automatically delineate the liver, spleen, left and right autochthon muscles, and body regions. Skeletal muscle, subcutaneous fat, and visceral fat were automatically segmented using the method described in [9]. To generate kidney lesion labels, an nnU-Net model was trained on the full KiTS23 dataset (489 CT scans) and subsequently applied to the remaining four datasets listed in Table 1. Kidney cortex and medulla labels in the training cohort were manually annotated or refined through an iterative-learning workflow by a medical imaging researcher. For the testing cohort, all cortex and medulla labels were manually annotated by the same researcher. To ensure label quality, a subset of the annotations was independently reviewed by a radiologist with more than 30 years of experience. The resulting annotations were combined to create 13 class segmentation labels for both the training and testing datasets. The Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (HD95), and Average Surface Distance (ASD) were used to quantitatively assess cortex and medulla segmentation accuracy.

2.2 Stratified Iterative Learning for Kidney Cortex and Medulla Segmentation

The stratified iterative learning framework consisted of four iterations. In the first iteration, an initial nnU-Net model was trained using 10 normal CT scans

from the Sarcopenia dataset (Table 1). The remaining 64 CT scans without kidney lesions from the same dataset were then automatically segmented by the initial model, and the predicted segmentations were manually refined to produce a total of 74 labeled CT scans. These annotated scans were used to train an improved nnU-Net model in the second iteration. In the third iteration, the same workflow was applied, with the addition of 22 CT scans containing kidney lesions. These scans were first segmented by the model from the second iteration and then manually refined. Of the 22 lesion cases, 10 were the remaining scans from the Sarcopenia dataset that had not been used in the previous iterations, and the other 12 were obtained from the Prostate dataset. In the final iteration, an additional 50 CT scans with kidney lesions from the KiTS23 dataset were incorporated using the same iterative learning strategy, resulting in the final kidney cortex and medulla segmentation model.

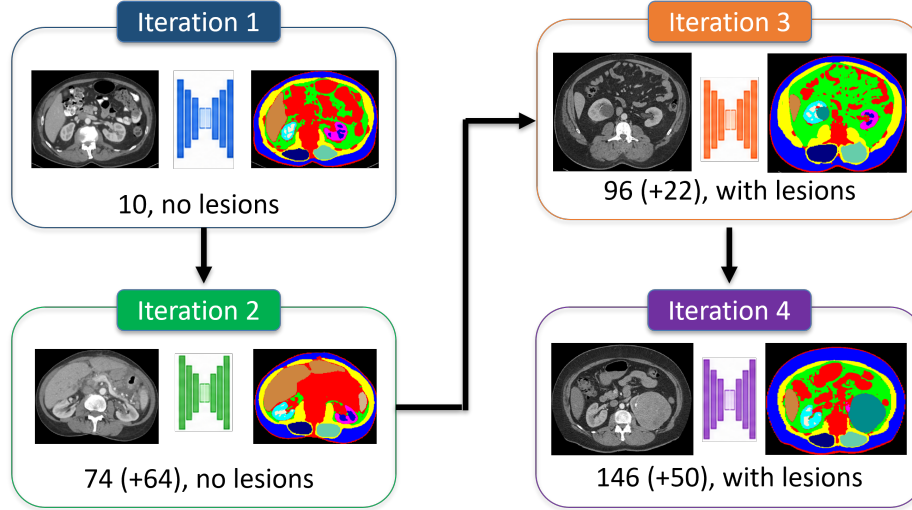


Fig. 1: Stratified iterative learning framework for kidney cortex and medulla segmentation, beginning with 10 normal CT scans and gradually adding lesion scans to minimize manual annotation.

All four segmentation models were developed using the nnU-Net framework [6]. The default nnU-Net training settings were adopted, including 1,000 epochs, a composite loss function consisting of binary cross-entropy and soft Dice loss with class-adaptive weighting, and the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 1×10^{-3} . Because the kidneys occupy fixed anatomical positions, with the left kidney always located on the left side of the body and the right kidney on the right side, mirror data augmentation was disabled during training. Applying left-right flipping would generate anatomically unrealistic CT volumes that do not occur in clinical practice by swapping the

positions of the two kidneys. Training with such artificial samples could confuse the network and increase the likelihood of misidentifying the left kidney as the right kidney, and vice versa, thereby reducing the anatomical consistency and accuracy of kidney cortex and medulla segmentation.

3 Experimental Results

Table 2 summarizes the segmentation accuracy on the Renal Hypertension and SAROS test datasets using the segmentation models trained in the four learning iterations. On the Renal Hypertension dataset, incorporating CT scans with kidney lesions substantially improved segmentation accuracy compared with models trained only on CT scans without kidney lesions. Adding more CT scans with kidney lesions further yielded a modest improvement in segmentation accuracy. In contrast, incorporating CT scans with kidney lesions had only a minimal impact on segmentation accuracy on the SAROS dataset. Instead, increasing the number of training CT scans without kidney lesions played a major role in improving performance, as evidenced by the notable accuracy gain from Iteration 1 to Iteration 2. Thereafter, the segmentation accuracy remained stable, with only slight fluctuations across the subsequent iterations.

Table 2: Dice similarity coefficient (DSC) of the combined left and right kidney cortex and medulla across the four learning iterations on the Renal Hypertension and SAROS test datasets.

Iteration #	1	2	3*	4*
Renal Hypertension	87.8 ± 16.1	87.8 ± 16.1	96.3 ± 2.7	97.0 ± 3.0
SAROS	94.9 ± 4.7	96.6 ± 3.6	96.5 ± 4.0	96.3 ± 2.7

*CT scans with kidney lesions were used to train the segmentation model in this iteration.

An example of kidney cortex and medulla segmentation in the presence of a kidney lesion from the SAROS dataset is shown in Fig. 2. The segmentation models from Iterations 1 and 2 failed to identify the lesion region because no CT scans with kidney lesions were included in the training data. After CT scans with kidney lesions were incorporated into the training set, the lesion region was included in the segmentation result in Iteration 3 but was incorrectly classified as the right kidney cortex. In Iteration 4, the lesion region was not only included in the segmentation result but also correctly identified as a kidney lesion. Overall, the segmentation accuracy progressively improved across the four learning iterations.

Table 3 presents the segmentation accuracy for the left kidney cortex, left kidney medulla, right kidney cortex, and right kidney medulla on the Renal Hypertension and SAROS test datasets. In both datasets, the kidney cortex

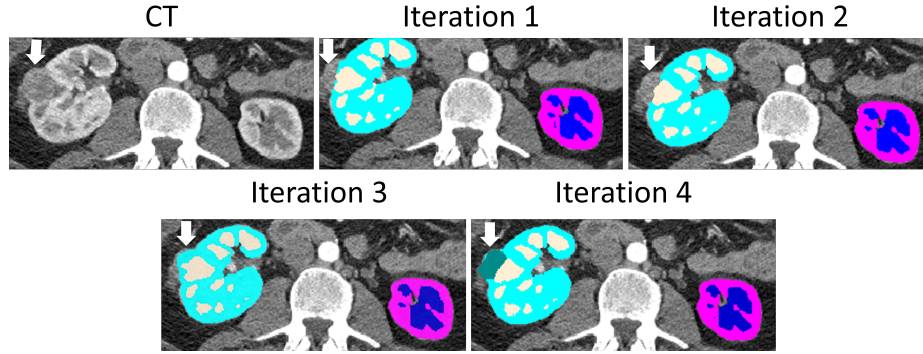


Fig. 2: Comparison of the kidney cortex and medulla segmentation results for the same patient using the segmentation models from the four learning iterations. The white arrows in all subfigures indicate a kidney lesion. Color coding: right kidney cortex, cyan; right kidney medulla, light brown; left kidney cortex, magenta; left kidney medulla, blue; and kidney lesion, dark green.

was segmented more accurately than the kidney medulla. The segmentation accuracies of the left and right kidney cortex were nearly identical, with a DSC difference of only 0.1% in the SAROS dataset. In contrast, greater variation was observed for kidney medulla segmentation, with the DSC difference between the left and right kidneys exceeding 1% on the SAROS dataset. Nevertheless, for all four anatomical structures, the DSC exceeded 94.7%, while the HD95 and ASD remained below 1.0 mm and 0.32 mm, respectively, even in the presence of kidney lesions.

Table 3: Segmentation accuracy on the Renal Hypertension and SAROS test datasets. Values are reported as mean $\pm$ standard deviation. $\uparrow$ indicates higher is better and $\downarrow$ indicates lower is better.

Dataset	Structure	DSC (%) $\uparrow$	HD95 (mm) $\downarrow$	ASD (mm) $\downarrow$
Renal Hypertension	Left cortex	98.1 ± 1.4	0.73 ± 0.28	0.32 ± 0.64
	Left medulla	95.6 ± 5.8	0.82 ± 0.62	0.14 ± 0.17
	Right cortex	98.1 ± 1.1	0.83 ± 0.11	0.07 ± 0.03
	Right medulla	96.2 ± 3.0	0.87 ± 0.42	0.10 ± 0.08
SAROS	Left cortex	97.3 ± 1.4	0.31 ± 0.39	0.08 ± 0.08
	Left medulla	95.9 ± 2.5	0.65 ± 1.17	0.11 ± 0.15
	Right cortex	97.2 ± 2.2	0.60 ± 1.42	0.09 ± 0.18
	Right medulla	94.7 ± 3.6	1.00 ± 1.30	0.14 ± 0.17

Fig. 3 presents representative kidney cortex and medulla segmentation results for four patients. Compared with the manual annotations shown in the middle column, the segmentations closely match the ground truth for both the patient without kidney lesions (top row) and the patients with kidney lesions of increasing size (second through fourth rows). These results demonstrate the high segmentation accuracy of the proposed method and its robustness to varying kidney lesion sizes.

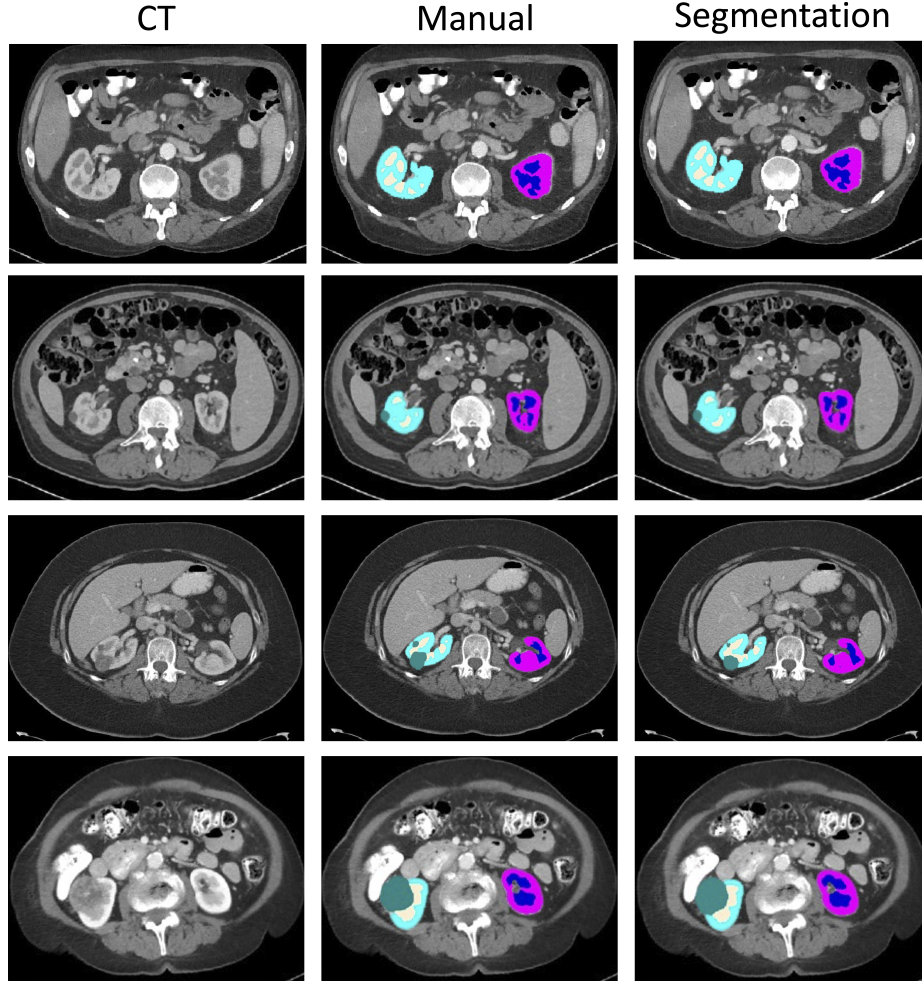


Fig. 3: Examples of kidney cortex and medulla segmentation for four patients, with one patient shown in each row. The patient in the first row has no kidney lesions, whereas the patients in the second through fourth rows exhibit kidney lesions of progressively increasing sizes.

4 Conclusion

Kidney cortex and medulla volume measurements have the potential to serve as important biomarkers for compartment-specific assessment of kidney diseases, including chronic kidney disease, diabetes, and renal fibrosis. In this work, we proposed a stratified iterative learning framework to develop an accurate kidney cortex and medulla segmentation model while facilitating the annotation of increasingly challenging pathological cases. The iterative learning process began with CT scans without kidney lesions to rapidly establish an initial segmentation model. This model was then used to generate preliminary segmentations for newly added CT scans, requiring only manual refinement to produce high-quality annotations. More challenging CT scans with kidney lesions were introduced in later iterations, allowing the progressively improved models to provide increasingly accurate initial segmentations for subsequent manual refinement.

Experimental results demonstrated high segmentation accuracy. The DSC exceeded 97.2% for kidney cortex segmentation and 94.7% for kidney medulla segmentation, even in the presence of kidney lesions. Furthermore, the segmentation accuracy achieved in the early learning iterations (Table 2) indicates that only limited manual refinement was required in subsequent iterations, as the earlier models already produced accurate initial segmentations. Future work will focus on substantially expanding the evaluation dataset to include more challenging cases, particularly patients with multiple or large kidney lesions that significantly deform the kidney anatomy. Such validation will further assess the robustness and generalizability of the proposed segmentation framework in clinically complex scenarios. Future studies will also include independent annotations from multiple experts and assess inter-annotator agreement to provide a more rigorous evaluation of segmentation performance.

Prospect of application: The proposed kidney cortex and medulla segmentation framework can support automated volumetric analysis for compartment-specific assessment of kidney diseases, including chronic kidney disease, diabetes, and renal fibrosis. It can be integrated into clinical and research workflows to provide rapid, reproducible kidney volume measurements from routine CT scans.

Acknowledgments. This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center and the National Institute of Allergy and Infectious Diseases. This work utilized the computational resources of the NIH HPC Biowulf cluster (<https://hpc.nih.gov>).

Disclaimer This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

Disclosure of Interests. Ronald M. Summers has received royalties from iCAD, Philips, and ScanMed. All other authors declare no competing interests relevant to the content of this article.

References

1. Beland, M.D., Walle, N.L., Machan, J.T., Cronan, J.J.: Renal Cortical Thickness Measured at Ultrasound: Is It Better Than Renal Length as an Indicator of Renal Function in Chronic Kidney Disease? *American Journal of Roentgenology* **195**(2), W146–W149 (2010)
2. Van den Broucke, S., et al.: Segmentation-based quantitative measurements in renal CT imaging using deep learning. *European Radiology Experimental* **8**, 110 (2024)
3. Evans, R.G., Eppel, G.A., Anderson, W.P., Denton, K.M.: Mechanisms Underlying the Differential Control of Blood Flow in the Renal Medulla and Cortex. *Journal of Hypertension* **22**(8), 1439–1451 (2004)
4. Heller, N., Wood, A., Isensee, F., Rädtsch, T., Tejpaul, R., Papanikolopoulos, N., Weight, C. (eds.): *Kidney and Kidney Tumor Segmentation: MICCAI 2023 Challenge, KiTS 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings, Lecture Notes in Computer Science*, vol. 14540. Springer, Cham, Switzerland (2024)
5. Humphreys, B.D.: Mechanisms of Renal Fibrosis. *Annual Review of Physiology* **80**, 309–326 (2018)
6. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203–211 (2020)
7. Koitka, S., Baldini, G., Kroll, L., van Landeghem, N., Pollok, O.B., Haubold, J., Pelka, O., Kim, M., Kleesiek, J., Nensa, F., Hosch, R.: Saros: A dataset for whole-body region and organ segmentation in ct imaging. *Scientific Data* **11**, 483 (2024)
8. Korfiatis, P., Denic, A., Edwards, M.E., Gregory, A.V., Wright, D.E., Mullan, A., Augustine, J., Rule, A.D., Kline, T.L.: Automated Segmentation of Kidney Cortex and Medulla in CT Images: A Multisite Evaluation Study. *Tomography* **10**(5), 2697–2712 (2024)
9. Liu, J., Bhadra, S., Shafaat, O., Mukherjee, P., Parnell, C., Summers, R.M.: A unified approach to medical image segmentation by leveraging mixed supervision and self and transfer learning (mist). *Computerized Medical Imaging and Graphics* **122**, 102517 (2025)
10. Luo, H., Li, J., Huang, H., Jiao, L., Zheng, S., Ying, Y., Li, Q.: AI-based segmentation of renal enhanced CT images for quantitative evaluation of chronic kidney disease. *Scientific Reports* **14**, 16890 (2024)
11. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence* **5**(5), e230024 (2023)
12. Xiang, D., Bagci, U., Jin, C., Shi, F., Zhu, W., Yao, J., Sonka, M., Chen, X.: CorteXpert: A Model-Based Method for Automatic Renal Cortex Segmentation. *Medical Image Analysis* **42**, 257–273 (2017)
13. Zeni, L., Norden, A.G.W., Cancarini, G., Unwin, R.J.: A More Tubulocentric View of Diabetic Kidney Disease. *Journal of Nephrology* **30**(6), 701–717 (2017)