Automatic segmentation of polycystic kidneys from magnetic resonance images using a three-dimensional fully-convolutional network

Jamie A. O'Reilly[1,*], Sakuntala Tanpradit[1], Tasawan Puttasakul[1], Manas Sangworasil[1], Takenobu Matsuura[1], Khaisang Chousangsuntorn[2] and Pornphan Wibulpolprasert[3]

[1]College of Biomedical Engineering, Rangsit University, Pathum Thani, Thailand
[2]Department of Radiological Technology, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand
[3]Department of Diagnostic and Therapeutic Radiology, Faculty of Medicine, Mahidol University, Bangkok, Thailand
[*]Corresponding author, E-mail: jamie.o@rsu.ac.th

Abstract

Autosomal dominant polycystic kidney disease (ADPKD) is a genetic disorder that causes progressive cyst development in the kidneys. This enlarges the kidneys, disturbing glomerular filtration, ultimately producing kidney failure. The process of cyst growth leads to severely atypical morphology. Combined with the presence of extrarenal cysts, this makes identifying the kidneys from medical images challenging; requiring the expertise of trained radiologists. This is time-consuming and suffers from inter- and intra-observer variability; hence our objective is to automate the segmentation of kidneys from ADPKD patient scans. A total of 135 T2-weighted MRI scans were obtained from 55 patients aged 16 to 87 years (mean = 55.94 ± 13.39 SD), weighing 35 to 90 kg (59.39 ± 11.16), with a sex distribution of 29:106 (male: female), and total kidney volume (TKV) for left or right kidneys ranging from 83.77 to 3,376.78 ml (633.85 ± 520.14). These were all annotated with kidney labels and split into training (n = 120) and test (n = 15) sets for supervised learning with a 3D fully-convolutional network model (i.e. "3D-Unet"/"V-net"). Accuracy and dice similarity coefficient (DSC) were used to evaluate segmentation performance, and coefficient of determination (R^2) was used to compare automatically-derived TKV with clinical reference measurements. The results of nine-fold cross validation demonstrated test set accuracy of 97.72 ± 0.8 %, DSC of 0.787 ± 0.060. Comparison of TKV measurements showed R^2 of 0.787 after removing ten edge cases. These results are encouraging and indicate the potential of this technology, although further development and careful evaluation are needed before clinical adoption.

Keywords: *Segmentation, Deep Learning, Polycystic Kidneys, Fully Convolutional Network, Medical Image Processing, Machine Learning, TKV*

1. Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is an inherited disorder characterized by the formation of multiple renal cysts, which can also be accompanied by extrarenal cysts, mainly inside of the liver or other peritoneal tissues (Pei, 2006; Pirson, 2010; Torra et al., 2008). Due to considerable heterogeneity in manifestation of ADPKD severity in families and between families, it is thought that a gene-environment interaction contributes to the development of this disease (Torres, Rossetti, & Harris, 2007). With aging, patients typically exhibit increasing numbers of cysts that also tend to grow in size. The cyst load adds to the total kidney volume (TKV), stressing and deforming the renal parenchyma. Enlargement and distortion of the kidneys eventually disrupts the physiological processes required for normal glomerular filtration (Grantham et al., 2006). In consequence, this can lead to end-stage renal disease, which can be fatal if kidney transplantation is not performed. The chronic nature of ADPKD is rather stark for sufferers, and there are few effective treatments available. Disease management generally involves trying to maintain a healthy lifestyle with recommended fluid intake, and having regular check-ups that involve performing a magnetic resonance imaging (MRI) or computed tomography (CT) scan to determine TKV (Wong et al., 2018). Of these two modalities, for regular (two or more times per year) imaging MRI is preferable due to the increased health risks associated with exposure to radiation. Medical imaging procedures are performed after which the radiologist must analyze each image to determine the boundary of kidney tissue. This process may be referred to as manual segmentation and is a standard practice in the management of ADPKD. Radiologists typically perform manual segmentation with the help of a digital drawing tool. Currently,

although manual segmentation is the best available method for segmenting the kidneys, it has several drawbacks; it is time-consuming and suffers from inter- and intra-observer variability. The time of specialist healthcare professionals is precious, and routinely spending in performing laborious image analysis is inefficient. Furthermore, variability between observers, or by the same observer under different conditions, is also undesirable. An automatic computational approach to segmenting the kidneys could unburden clinicians from the tedious elements of this task. High precision and repeatability are clearly required from any computational method that may be seriously considered for introduction to clinical practice. Following kidney segmentation, TKV can be calculated, which is used to track disease progression (Alam et al., 2015).

There have been researched efforts dedicated towards this problem since the turn of the century (Zöllner et al., 2012). The development of computational methods for automatically segmenting polycystic kidneys may be classified separately into conventional algorithm-based and data-driven machine learning or deep-learning model-based approaches. It has become widely recognized that deep-learning model-based methods using fully-convolutional network (FCN) architectures tend to produce better results. For example, using this approach, Sharma et al. (2017) achieved a DSC of 0.86 with a total of 244 CT scans, which they deemed to be in agreement with clinical experts; meanwhile, Kline et al. (2017) managed to achieve a DSC of 0.96 with 2400 MRI scans of the kidneys. Anecdotally this exemplifies the relationship between the amount of available training data and performance for deep-learning approaches. Recently, segmentation accuracy of 90 % was reported based on a two-model classification-segmentation pipeline trained and tested with 526 MR images from 18 patients (Brunetti et al., 2019). Taken together, this prior work suggests that automatic segmentation of polycystic kidneys from MRI scans of ADPKD patients using FCN models is feasible with varying amounts of data. Differences in scan format and patient demography may limit the successful transfer of models trained on data from different distributions; thus we aim to develop a model for a specific clinic.

The three-dimensional FCN architecture for medical volume segmentation was introduced relatively recently (Çiçek, Abdulkadir, Lienkamp, Brox, & Ronneberger, 2016; Milletari, Navab, & Ahmadi, 2016). This model typically begins with an encoding path comprising multiple convolutional and pooling layers, typically discarding some portion of units at random to mitigate overfitting (Srivastava, Hinton, Krizhevsky, & Salakhutdinov, 2014), and normalizing layer inputs to make the model more robust to learning rate and initialization parameters (Ioffe & Szegedy, 2015). During the encoding part of the FCN, the model learns increasingly abstract information at the expense of spatial resolution. The encoding structure feeds into a decoding section comprising the same number of up-sampling transposed convolutional layers, with "skip connections" cascading equal resolution layers from the encoding path to preserve spatial information. The model used in the present study is illustrated in Figure 1. Several variants of this FCN architecture have been proposed, some of which include recurrent and/or residual network topologies; although there is no consensus about what combination of layers results in a superior model, and typically model ensembles are found to produce the best overall performance (Heller et al., 2019; Isensee & Maier-Hein, 2019).

This FCN architecture has been quite successful in a range of biomedical image segmentation problems (Guo, Guo, Li, & Gong, 2019; Ye, Wang, Zhang, & Wang, 2019; Zhang et al., 2019; Zhong et al., 2018), and has become particularly prominent among submissions to Grand Challenges in Biomedical Image Analysis (https://grand-challenge.org/). Pertinent among these competitions is the kidney and kidney tumor segmentation challenge (KiTS19; Heller et al., 2019), in which the top eight scoring submissions employed variants of the 3D FCN architecture. Albeit this competition used contrast-enhanced CT image data, the results are nevertheless encouraging for the prospects of developing computational methods for automatic kidney segmentation at comparable levels to expertly trained humans. At present, it appears as though the 3D FCN architecture has become the *de facto* standard method for biomedical volume segmentation. As mentioned, this approach certainly has achieved a lot of acclaim for its performance in various articles and publicized challenges; although each new application with specific data requirements must be carefully validated because there are no guarantees that this approach will be equally successful in all applications.

2. Objectives

In this study, we aim to explore the application of a 3D FCN architecture for automatic segmentation of polycystic kidneys from MRI scans of ADPKD patients. It should be noted that this objective relates only to data in the format described in Section 3.1, and strictly does not apply to different medical imaging modalities (e.g. CT), scanning machines, specific settings, or demographic groups; all of which may influence the performance of data-driven methodologies.

3. Materials and Methods

3.1. Data

The data used in this study consisted of coronal single-shot T2-weighted MRI scans (n = 135) from 55 patients diagnosed with ADPKD who underwent kidney volume assessment between January 1, 2014, and September 24, 2019, at Ramathibodi Hospital, Bangkok, Thailand. This cohort ranged in age from 16 to 87 years (55.94 ± 13.39), in weight from 35 to 90 kg (59.39 ± 11.16), and sex ratio of 29:106 (*male*: *female*). Total kidney volume measurements were calculated from manual segmentations and ranged from 83.77 to 3,376.78 ml (633.85 ± 520.14). Kidney volume distributions are shown in Figure 1. Scan dimensions were square in the coronal plane with side length ranging from 384 to 672 pixels (501.33 ± 39.82) and pixel spacing from 0.586 to 0.804 mm (0.706 ± 0.030). The number of frames ranged from 26 to 44 (33.68 ± 4.33) with a constant frame spacing of 5 mm. These differences in the scan dimension reflect body size variability within the patient group.

Binary masks corresponding to the kidneys were created manually to facilitate the supervised learning approach by providing labels for every voxel. Imaging data were normalized to values between 0 and 255. Due to hardware limitations and resource requirements of the model, volume data were resized to 16 x 64 x 64 (*frames* x *rows* x *columns*). One hundred and twenty scans were used for model training and fifteen were used for testing; for nine-fold cross-validation different portions of data were extracted for training and testing. Throughout this manuscript values are reported as *mean* ± *standard deviation*, unless otherwise stated.

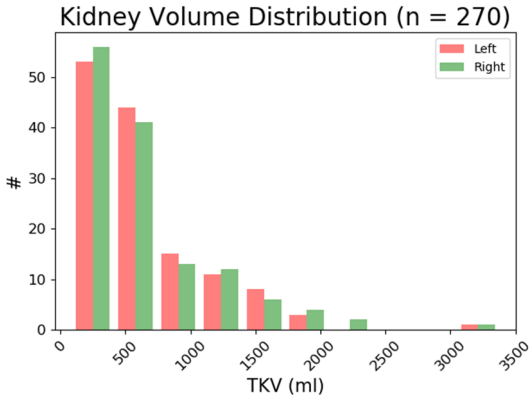

Figure 1 Distribution of kidney volume within the dataset. In total there were 270 kidneys from 135 patients.

3.2. Model Architecture

The 3D FCN model for volume segmentation was based on those previously mentioned (Çiçek et al., 2016; Milletari et al., 2016). The encoding section included five levels, each with 3 x 3 x 3 convolutional layers followed by 2 x 2 x 2 max pooling with a stride of two. Every pooling layer of the encoding path was followed by batch normalization, and layers two and four were followed by 50% dropout. The decoding section of the model consisted of five levels concatenated with layers of equal size in the encoding section, combined with convolutional and up-sampling layers. This combination of encoding and decoding sections accepted an input volume size of 16 x 64 x 64 and the output volume of the same size, as illustrated in Figure

Proceedings of RSU Internationl Research Conference (2020)

2. Input and hidden layers used the rectified linear unit (ReLU) activation function, while the sigmoid function was applied at the output layer.

The model was trained for 1000 iterations with adaptive momentum (Adam) optimization, an initial learning rate set as 1 x 10^{-4}, β1 = 0.9, and β2 = 0.99. Weighted binary cross-entropy loss was incorporated to account for imbalanced classes, with class weights calculated using the training set. Training data was augmented during this process by randomly flipping axes, shifting (up to 50 pixels left or right) and rotating (up to 45° clockwise or anticlockwise) in the coronal plane, and adjusting pixel intensities (up to 50 points brighter or darker), before resizing to comply with the model input size.

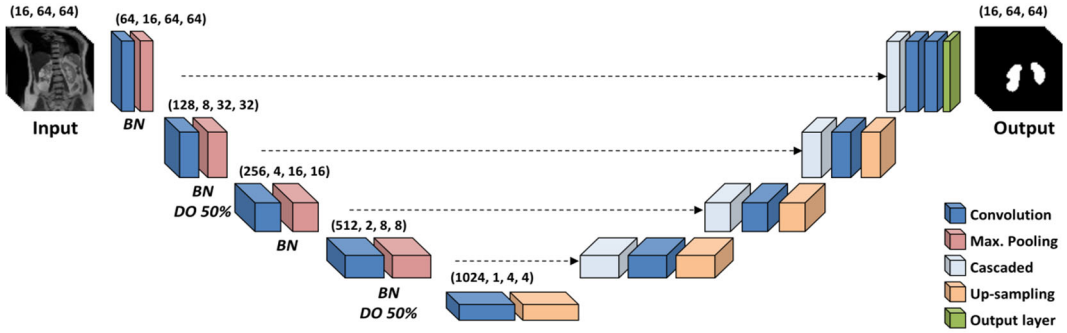

Figure 2 Fully-convolutional network architecture. This diagram illustrates the "U-net" topology. Application of batch normalization (*BN*) and dropout (*DO*) are annotated below the relevant levels of the encoder section; data sizes at each level of resolution are denoted in parentheses; skip connections to concatenate layers are represented by dashed arrows.

3.3. Evaluation

Nine-fold cross validation was performed to evaluate the model over the whole dataset; in each fold test data was unseen by the model. Segmentation performance was quantified using accuracy and Dice similarity coefficient (DSC) metrics. These provided end-point assessments (Table 2) and longitudinal evaluations throughout the training process (Figure 3). Furthermore, to determine correspondences between TKV values derived from automatic segmentations and clinical reference measurements, coefficient of determination (R^2), mean absolute error (MAE) and mean absolute percentage error (MAPE) were calculated. These evaluation metrics were calculated as shown in Table 1.

Table 1 Evaluation metrics

Metric	Evaluation	Formula
Accuracy	Segmentation	$\frac{TP+TN}{TP+FP+TN+FN}$
DSC	Segmentation	$\frac{2TP}{2TP+FP+FN}$
R^2	TKV	$1-\frac{\sum(Ref_i-Auto_i)^2}{\sum(Ref_i-\overline{Ref})^2}$
MAE	TKV	$\frac{1}{n}\sum_{i=0}^{n}\lvert Ref_i-Auto_i\rvert$
MAPE	TKV	$\frac{1}{n}\sum_{i=0}^{n}\left\lvert\frac{Ref_i-Auto_i}{Ref_i}\right\rvert\times 100\%$

TP = true positive; FP = false positive; TN = true negative; FN = false negative; n = number of test cases; $Auto_i$ = TKV from auto-segmentation; Ref_i = TKV reference measurement; $\overline{Ref}$ = average from test case reference TKV measurements.

3.4. Tools

Software used in this study included Scikit-Learn 0.20.2, OpenCV 4.0.0.21, Keras 2.2.4, and Tensorflow 1.13.1 for Python 3.7.2. Hardware included an Intel i7-9700 CPU with 32 GB RAM and an NVIDIA GeForce GTX 1070 8 GB graphical processing unit.

3.5. Ethics

This study was approved by the Ethics Committee on Human Rights Related to Research Involving Human Subjects of Ramathibodi Hospital. Anonymized data with no identifiable information was used.

4. Results and Discussion

Accuracy and DSC from nine-fold cross-validation are reported in Table 2. These equate to an overall test set accuracy of 97.72 ± 0.827 % and DSC of 0.787 ± 0.060. Total kidney volume for left and right kidneys were estimated based on these segmentations (Figure 3); however, it should be noted that FCN output did not differentiate between kidneys, and binary segmentation masks were split post-hoc depending on the position of respective centers of mass. In six instances a single large kidney region was segmented, due to their large and abnormal anatomy, in which cases the total volume was halved to estimate left and right kidney volumes. Two cases only segmented one kidney, and a further two cases detected no kidney at all. Analysis of TKV from the whole dataset shown in Figure 3a was evaluated to have R^2 of 0.617, MAE of 187.5 ml, and MAPE of 36.71 %. After removing these ten edge cases (Figure 3b), the performance was re-evaluated to be R^2 of 0.787, MAE of 147.1 ml, and MAPE of 35.34 %. This suggests that there are improvements to be made in terms of segmentation splitting algorithm.

Table 2 Segmentation performance evaluation

k	Test Set Accuracy	Test Set DSC
1	0.983	0.866
2	0.960	0.824
3	0.985	0.802
4	0.983	0.683
5	0.978	0.780
6	0.985	0.843
7	0.979	0.834
8	0.966	0.698
9	0.976	0.757

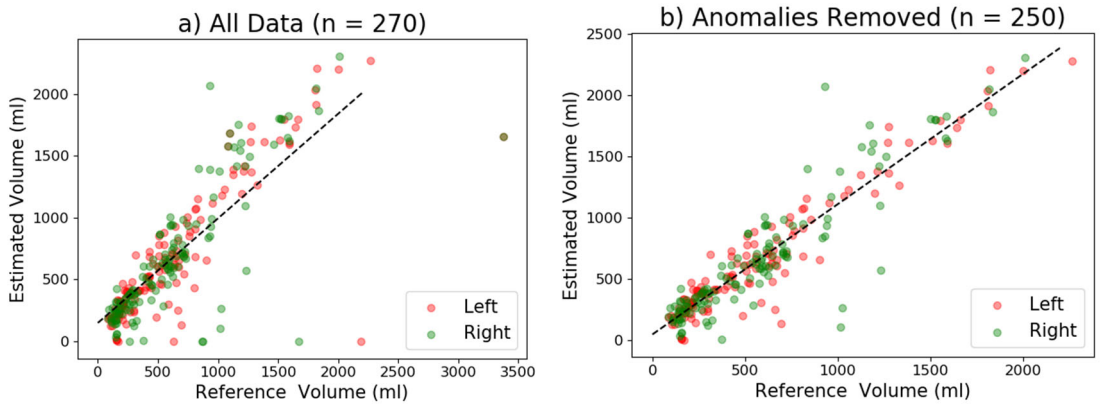

Figure 3 Comparisons of kidney volumes from manual and automatic segmentations. a) Whole dataset. b) Ten anomalous cases have been removed. Reference volumes were derived from manual segmentations; estimated volumes were computed from automatic segmentations.

The learning curves in figure 4 describe model performance throughout the training process. In terms of accuracy, the model exhibits low bias and variance. However, in terms of DSC, model bias and variance are greater; DSC is a more stringent metric because it does not count true negative voxel classifications, hence it is generally considered to be more appropriate for evaluating minority class segmentations. Trajectories of the learning curves suggest that there is a need to improve generalizability; further developments shall seek to reduce the variance. Approaches to improving this performance may potentially include obtaining more training data and modifying the FCN model design.

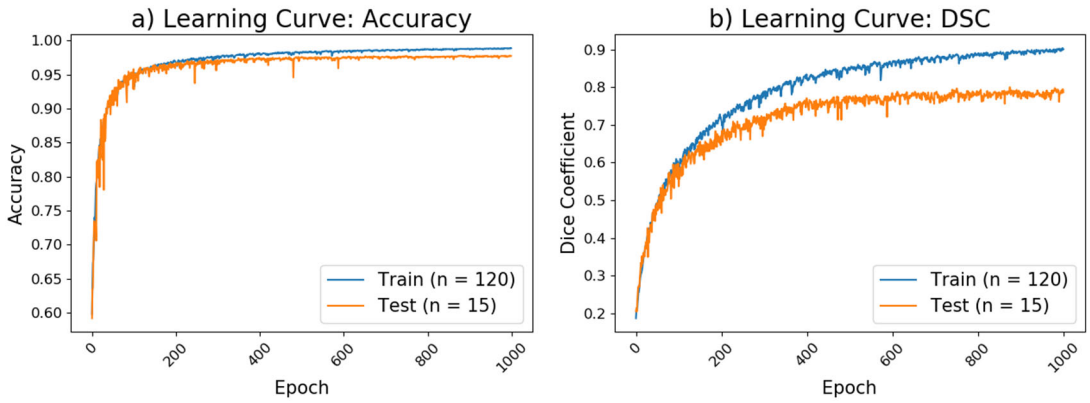

Figure 4 Learning curves. a) Accuracy. b) Dice similarity coefficient. Traces show the average from nine-fold cross-validation.

In comparison with previously published attempts at using FCN methods for automatically segmenting ADPKD patient kidneys from medical images, this study used a modestly sized dataset; i.e. 135 scans in the present study versus 244 (Sharma et al., 2017) and 2400 (Kline et al., 2017). These two previous studies reported DSCs of 0.86 and 0.96, respectively. The finding of 0.79 DSC from the present study may, therefore, reflect a lower amount of data used for model training. Furthermore, these two prior studies segmented the kidneys from medical imaging scans on a per image basis, which would have provided orders of magnitude more training instances (e.g. 244 CT scans ≈ 25,000 CT images). In order to achieve comparable performance, we may continue to collect more ADPKD patient scans in the anticipation that more data will equate to a better model, which is typically the case with deep learning technology.

Paradoxically, the seemingly reasonable performance of automatic segmentation (accuracy of 97.7 % and DSC of 0.79) does not translate directly into equivalently accurate TKV measurements (R^2 of 0.62 and MAPE of 36.71 %). This may be due to a loss of resolution by resizing the data for the FCN input layer; i.e. down-sizing patient scans from, for example, 32 x 480 x 480 (*frames* x *rows* x *columns*) to 16 x 64 x 64. It is expected that training a larger model with an input size closer to the original data size will produce more accurate final segmentations and corresponding TKV estimates. There are also issues regarding the kidney splitting algorithm used to separate large individual regions in a scan that needs to be addressed.

5. Conclusion

In conclusion, 3D FCN technology can potentially achieve suitable levels of performance for clinical adoption; however, the current method should be further developed before considering deployment. To improve precision we shall 1) use more advanced hardware to develop a higher resolution model, 2) collect more data on the basis that this will improve model generalization, and 3) experiment with different network architectures. We anticipate that this technology will be introduced to radiological practice in the foreseeable future.

6. Acknowledgements

We are grateful for technical assistance from Mr. Rawiphon Chotikunnan and access to computing resources from Dr. Pattarapong Phasukkit. Funding for this project was received from the Research Institute of Rangsit University (grant number 90/2561).

7. References

Alam, A., Dahl, N. K., Lipschutz, J. H., Rossetti, S., Smith, P., Sapir, D., … Bichet, D. G. (2015). Total kidney volume in autosomal dominant polycystic kidney disease: A biomarker of disease progression and therapeutic efficacy. *American Journal of Kidney Diseases*, *66*(4), 564–576. https://doi.org/10.1053/j.ajkd.2015.01.030

Brunetti, A., Cascarano, G. D., De Feudis, I., Moschetta, M., Gesualdo, L., & Bevilacqua, V. (2019). Detection and Segmentation of Kidneys from Magnetic Resonance Images in Patients with Autosomal Dominant Polycystic Kidney Disease. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11644 LNCS, pp. 639–650). https://doi.org/10.1007/978-3-030-26969-2_60

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9901 LNCS*, 424–432. https://doi.org/10.1007/978-3-319-46723-8_49

Grantham, J. J., Torres, V. E., Chapman, A. B., Guay-Woodford, L. M., Bae, K. T., King, B. F., … Miller, J. P. (2006). Volume progression in polycystic kidney disease. *New England Journal of Medicine*, *354*(20), 2122–2130. https://doi.org/10.1056/NEJMoa054341

Guo, Z., Guo, N., Li, Q., & Gong, K. (2019). Automatic multi-modality segmentation of gross tumor volume for head and neck cancer radiotherapy using 3D U-Net. In H. K. Hahn & K. Mori (Eds.), *Medical Imaging 2019: Computer-Aided Diagnosis* (p. 8). SPIE. https://doi.org/10.1117/12.2513229

Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., … Papanikolopoulos, N. (2019). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 Challenge. Retrieved from http://arxiv.org/abs/1912.01054

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015* (Vol. 1, pp. 448–456). International Machine Learning Society (IMLS).

Isensee, F., & Maier-Hein, K. H. (2019). An attempt at beating the 3D U-Net. University of Minnesota. https://doi.org/10.24926/548719.001

Kline, T. L., Korfiatis, P., Edwards, M. E., Blais, J. D., Czerwiec, F. S., Harris, P. C., … Erickson, B. J. (2017). Performance of an Artificial Multi-observer Deep Neural Network for Fully Automated Segmentation of Polycystic Kidneys. *Journal of Digital Imaging*, *30*(4), 442–448. https://doi.org/10.1007/s10278-017-9978-1

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016* (pp. 565–571). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/3DV.2016.79

Pei, Y. (2006, September 1). Diagnostic approach in autosomal dominant polycystic kidney disease. *Clinical Journal of the American Society of Nephrology : CJASN*. American Society of Nephrology. https://doi.org/10.2215/CJN.02190606

Pirson, Y. (2010). Extrarenal Manifestations of Autosomal Dominant Polycystic Kidney Disease. *Advances in Chronic Kidney Disease*, *17*(2), 173–180. https://doi.org/10.1053/j.ackd.2010.01.003

Sharma, K., Rupprecht, C., Caroli, A., Aparicio, M. C., Remuzzi, A., Baust, M., & Navab, N. (2017). Automatic Segmentation of Kidneys using Deep Learning for Total Kidney Volume Quantification in Autosomal Dominant Polycystic Kidney Disease. *Scientific Reports*, *7*(1), 2049. https://doi.org/10.1038/s41598-017-01779-0

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. *Journal of Machine Learning Research* (Vol. 15).

Torra, R., Sarquella, J., Calabia, J., Martí, J., Ars, E., Fernández-Llama, P., & Ballarin, J. (2008). Prevalence of cysts in seminal tract and abnormal semen parameters in patients with autosomal dominant polycystic kidney disease. *Clinical Journal of the American Society of Nephrology : CJASN*, *3*(3), 790–793. https://doi.org/10.2215/CJN.05311107

Torres, V. E., Rossetti, S., & Harris, P. C. (2007, April 14). Update on autosomal dominant polycystic kidney disease. *Minerva Medica*. Elsevier. https://doi.org/10.1016/S0140-6736(07)60601-1

Wong, A. T. Y., Mannix, C., Grantham, J. J., Allman-Farinelli, M., Badve, S. V., Boudville, N., … Rangan, G. K. (2018). Randomised controlled trial to determine the efficacy and safety of prescribed water intake to prevent kidney failure due to autosomal dominant polycystic kidney disease (PREVENT-ADPKD). *BMJ Open*, *8*(1). https://doi.org/10.1136/bmjopen-2017-018794

Ye, C., Wang, W., Zhang, S., & Wang, K. (2019). Multi-depth fusion network for whole-heart CT image segmentation. *IEEE Access*, *7*, 23421–23429. https://doi.org/10.1109/ACCESS.2019.2899635

Zhang, L., Luo, Z., Chai, R., Arefan, D., Sumkin, J., & Wu, S. (2019). Deep-learning method for tumor segmentation in breast DCE-MRI. In P. R. Bak & P.-H. Chen (Eds.), *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications* (p. 14). SPIE. https://doi.org/10.1117/12.2513090

Zhong, Z., Kim, Y., Zhou, L., Plichta, K., Allen, B., Buatti, J., & Wu, X. (2018). 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In *Proceedings - International Symposium on Biomedical Imaging* (Vol. 2018-April, pp. 228–231). IEEE Computer Society. https://doi.org/10.1109/ISBI.2018.8363561

Zöllner, F. G., Svarstad, E., Munthe-Kaas, A. Z., Schad, L. R., Lundervold, A., & Rørvik, J. (2012). Assessment of Kidney Volumes From MRI: Acquisition and Segmentation Techniques. *American Journal of Roentgenology*, *199*(5), 1060–1069. https://doi.org/10.2214/AJR.12.8657