Automated Assessment of Renal Calculi in Serial Computed Tomography Scans

Pritam Mukherjee[1](✉), Sungwon Lee[1], Perry J. Pickhardt[2], and Ronald M. Summers[1]

[1] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD, USA
pritam.mukhrejee@nih.gov

[2] Department of Radiology, School of Medicine and Public Health, The University of Wisconsin, Madison, WI, USA

Abstract. An automated pipeline is developed for the serial assessment of renal calculi using computed tomography (CT) scans obtained at multiple time points. This retrospective study included 722 scans from 330 patients chosen from 8544 asymptomatic patients who underwent two or more CTC (CT colonography) or non-enhanced abdominal CT scans between 2004 and 2016 at a single medical center. A pre-trained deep learning (DL) model was used to segment the kidneys and the calculi on the CT scans at each time point. Based on the output of the DL, 330 patients were identified as having a stone candidate on at least one time point. Then, for every patient in this group, the kidneys from different time points were registered to each other, and the calculi present at multiple time points were matched to each other using proximity on the registered scans. The automated pipeline was validated by having a blinded radiologist assess the changes manually. New graph-based metrics are introduced in order to evaluate the performance of our pipeline. Our method shows high fidelity in tracking changes in renal calculi over multiple time points.

Keywords: Renal calculi · Serial assessment · Deep learning · Registration

1 Introduction

CT colonography (CTC) is a nonenhanced CT of the abdomen and pelvis for detecting colorectal polyps but can be used for the opportunistic screening for kidney stones. Indeed, CT can be considered the diagnostic "gold standard", with accuracy close to 100% due to the higher attenuation of renal calculi compared to the surrounding tissue [1, 2]. Studies [3–5] have found the presence of asymptomatic kidney stones in 5–8% of CTC scans. Though considered a less significant finding, patients with asymptomatic kidney stones are known to be at increased risk of future symptomatic stone events [6] as more than half of the stones are reported to grow on longitudinal follow-up [7]. Fortunately, once stones are detected, suitable medications and dietary interventions have proven very effective, with dramatic reductions of more than 50% in recurrence rates [8]. Given the high clinical and financial burden of kidney stones (and urinary

S. Wu et al. (Eds.): AMAI 2022, LNCS 13540, pp. 39–48, 2022.
https://doi.org/10.1007/978-3-031-17721-7_5

stones, in general) with nearly 200,000 hospitalizations and an estimated annual cost exceeding $2 billion in 2000 [9], opportunistic detection and tracking of kidney stones can be vital.

Unfortunately, kidney stones, especially small asymptomatic ones, are often not reported or measured [10, 11]. Even when they are measured, inter-reader variability due to various factors such as CT window level etc.[12], and movement of the stones in the kidney during the follow-up interval tends to make serial assessment of stones difficult, time consuming and unreliable. Despite the importance and widespread use of serial imaging in clinical decision making [13], few machine learning studies to date have focused on the analysis of serial scans. Here, we propose a fully automated computing pipeline to detect and track stones on CTC over multiple follow-up scans.

1.1 Our Contributions

While several kidney stone detectors (deep learning based or otherwise) have been presented in the literature, to the best of our knowledge, tracking and serial assessment of kidney stones has largely remained unexplored. This work is a first step in that direction. Our main thrust in this paper is the tracking of stones over multiple scans – we leverage an existing deep learning based kidney stone detector [14] for the detection and segmentation step. The tracking step uses computationally cheap affine registration of bounding boxes containing the kidney – this enables accurate kidney registration while avoiding unintended deformations that may occur with deformable registration due to the relative movement of organs in the vicinity of the kidney. We evaluate an integrated software pipeline that combines both stone detection and stone tracking components and show that it is quite accurate despite its conceptual simplicity.

We also introduce new graph-based metrics for evaluating the performance of stone tracking. Existing multiple object tracking (MOT) metrics like multiple object tracking precision (MOTP) and multiple object tracking accuracy (MOTA)[15] may be inadequate for our task. MOTP measures the tracker's ability to precisely estimate the objects position independent of its skill at correctly matching the objects; while this is an important metric in near-real-time tracking in videos, this does not apply in our offline case with only 2–3 "frames" (corresponding to the scan time points), where the error in localizing the stones is zero if the stone is detected on the scan. MOTA, on the other hand does apply – it accounts for the total number of misses, false positives and mismatches over all frames and averaged over the total number of objects over all frames – and we use it to evaluate our full pipeline. However, MOTA is a patient-level metric and does not adequately capture the performance at the level of individual stone trajectories. Therefore, we introduce new graph-based metrics that capture the fidelity of tracking individual stones across multiple time points.

2 Materials and Methods

2.1 Data

This Health Insurance Portability and Accountability Act (HIPAA) compliant retrospective study was approved by the Institutional Review Board, and the need for signed informed consent was waived. The initial cohort comprised asymptomatic patients who underwent CT colonography (CTC) screening or follow-up between 2004 and 2016 at a single medical center. We only included patients who had nonenhanced CT scans with slice thickness less than 2 mm. We excluded (a) patients with no follow-up scans, (b) patients who exhibited stone-related symptoms in the follow-up interval, and (c) patients who underwent stone intervention such as extracorporeal shock wave lithotripsy (ESWL) between scans. A description of the indications for the cohort is in [16].

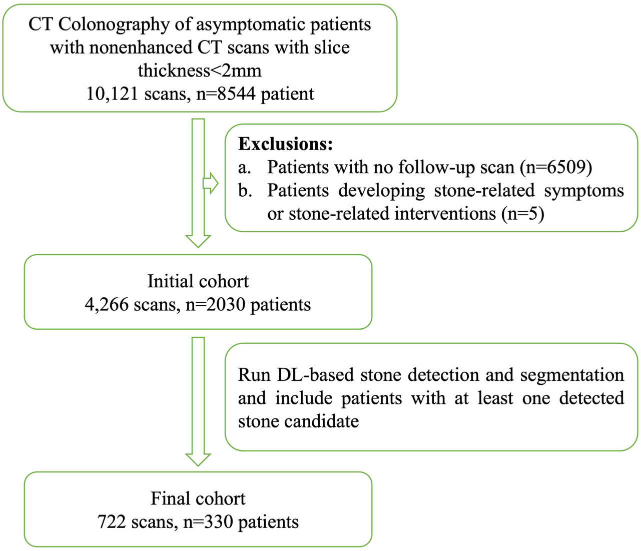
CT Colonography of asymptomatic patients with nonenhanced CT scans with slice thickness<2mm
10,121 scans, n=8544 patient
Exclusions:
a. Patients with no follow-up scan (n=6509)
b. Patients developing stone-related symptoms or stone-related interventions (n=5)
Initial cohort
4,266 scans, n=2030 patients
Run DL-based stone detection and segmentation and include patients with at least one detected stone candidate
Final cohort
722 scans, n=330 patients

Fig. 1. Cohort selection

2.2 Calculi Detection and Segmentation

The first step of the pipeline is the detection and segmentation of the calculi within the kidney. To that end, we use the model and method in [14] First, a 3D U-Net is used to segment the kidneys, followed by denoising and thresholding at the de facto standard 130 HU to identify candidate stones within the kidney. Finally, a CNN is used to predict if the candidate is truly a stone. The final output is a labeled volume with segmentations of each stone identified by the model. We use the pretrained model from [14] to identify and segment the stones in our dataset. The patients for which stones were detected by the model for at least one time point constituted our final cohort.

2.3 Registration and Stone Matching

For a given patient, and scans from two consecutive time points, we used affine registration to approximately align the kidneys and match the stones by location. The CT scans were first windowed (level: 50, width: 450) and normalized to the range [0, 1].

Registering the full abdominal CT scans is both computationally expensive and sensitive to the deformations of the large abdominal organs. We first extracted 3D “kidney boxes” – padded bounding boxes containing the kidney – from the CT scans. This was done by first labeling the kidney segmentation obtained from the model in the previous step using connected component analysis and assigning “left” and “right” designations based on their relative position in the axial images. Then the left and right kidney boxes from the first time point were individually registered to the left and right kidney boxes from the second time point. For the registration, we used the “greedy” registration software[17] (obtained from https://github.com/pyushkevich/greedy) with the following settings: 3D affine registration using the normalized cross-correlation metric with 4 × 4 × 4 patch size, with an initial transform matching image centers, performing a multi-resolution refinement using a maximum of 100 iterations at the coarsest level (4 ×), 50 iterations at an intermediate level (2 ×) and 10 iterations at full resolution (1 ×).

Once the individual boxes are registered, we perform stone matching in a greedy manner. First all stones from both time points are put in a bag. We then compute all pairwise Euclidean voxel distances $d\left(s_{1i}, s_{2j}\right)$, where s_{tk} refers to the k th stone at time point i. The stone pair with the minimum distance are “matched” and removed from the bag. We continue by matching the closest pair among the stones remaining in the bag, and so on until either the bag has stones from only one time point, or the minimum distance exceeds a predetermined distance threshold.

Overall, there are two hyperparameters in the stone tracking step: first, the padding (set at 0 by default) around the 3D bounding boxes in the registration step and the maximum distance threshold (set at $\sqrt{200}$ mm by default) in the stone matching step. We evaluate the effect of these two hyperparameters on the overall performance.

2.4 Manual Review and Tracking

The results of the stone detection and tracking were manually reviewed by a board-certified radiologist with 13 years of experience. The true stones were identified and tracked across the follow-ups.

2.5 Evaluation of Performance

We first evaluated the performance of the stone detection algorithm. Using the ground truth from manual review, we computed the precision and recall of the model.

Next, we evaluated the performance of registration and stone tracking. We did this in several ways. First, we used the widely used multiple object tracking accuracy (MOTA)[15] metric to evaluate the performance at the patient level. We also introduce several stone level performance metrics. We consider each stone (or candidate stone) identified in any scan as a vertex in a graph and add a directed edge (v_i, v_j) between two vertices v_i and v_j if they are determined to be the same stone from two time points i and j, and $i < j$. . We can thus construct a graph based on the ground truth stones and another on the predicted stones (see Fig. 2a for an example where both graphs have been merged). Using these graphs, we computed the following metrics:

(a) The precision and recall of the predicted edges. Heuristically, given the predicted edges (i.e., our pipeline determined that two stones from two consecutive time-points are the same stone), precision computes the fraction of them that are present in the ground-truth graph; conversely recall computes the fraction of edges in the ground-truth graph that are also present in the predicted graph. These metrics evaluate the tracking performance across two time points.

(b) The precision and recall of the predicted connected components. Each connected component represents a unique stone. Here, we are interested in evaluating how well we reproduce the full trajectory of the stones across all time points (possibly more than two). As before, precision computes the fraction of predicted stone trajectories that are present in the ground-truth graph, and recall computes the fraction of ground-truth stone trajectories that are present in the predicted graph. Note that a correctly predicted trajectory entails: (i) the stones being detected by the deep learning model in all time points, and (ii) the stones are correctly matched to each other across all consecutive time points. We also evaluate this metric for non-singleton connected components (stones or candidate stones that appear in the scans from only one time point).

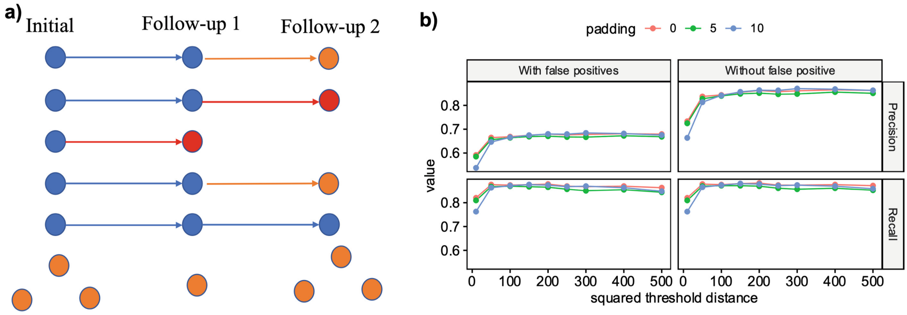

Fig. 2. (a.) Combined ground truth and predicted graphs for a patient. Each node is a stone. Blue indicates correct detection or tracking, orange indicates false positive predictions, red misses. (b.) Precision and recall of connected component retrieval with changing padding and threshold distance. It also shows the increase in precision when false positive stones are removed from the prediction graph before evaluating tracking. (Color figure online)

To disentangle the effect of stone detection and stone tracking, we repeat the above analysis by removing the false positive stone candidates. Further, after removing the false positive stone candidates, we consider two vertex-based metrics:

(a) The accuracy of retrieving a predecessor. Given a stone, we retrieve its predecessor from the previous time point. If the retrieved predecessor is the same for both the ground-truth graph and the predicted graph, we consider it correct. Stones from the initial scan necessarily do not have a predecessor; we remove these stones while computing this metric. However, new stones in follow-up scans may also lack predecessors – if both the ground-truth graph and the predicted graph agree, we consider it correct.

(b) The accuracy of retrieving a successor. Given a stone, we retrieve its successor from the next time point. If the retrieved successor is the same for both the ground-truth graph and the predicted graph, we consider it correct. Note that stones from the last follow-up scan do not have a successor; we remove these stones while computing this metric. However, stones that disappear in follow-up scans may also lack successors – if both the ground-truth graph and the predicted graph agree, we consider it correct.

2.6 Statistical Analysis

We use bootstrapping with 1000 iterates of patient sets randomly sampled with replacement from the original cohort to obtain confidence intervals, and the empirical quantiles at 2.5% and 97.5% are used as the 95% confidence intervals.

3 Results

3.1 Cohort Characteristics

Based on our inclusion criteria, 10,121 scans from 8544 patients were initially identified. After applying our exclusion criteria, we were left with 4266 nonenhanced CT scans from 2030 patients. The deep learning based kidney stone detector and segmenter [14] identified 837 stone candidates in 330 patients and 722 scans. These patients constituted our final cohort (see Fig. 1). A cohort characteristics summary is presented in Table 1.

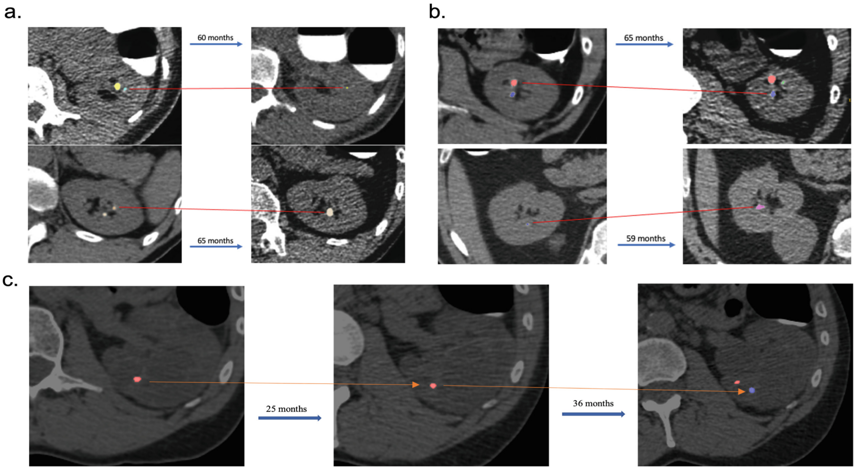

Fig. 3. Examples of registration. a. Inconclusive ground truth b. Incorrect stone matching. The red line shows the predicted matching. c. Correct matching over 3 time points

3.2 Performance of the Stone Detection and Segmentation

We evaluate the performance of the stone detection and segmentation based on the manual review by the radiologist. Out of 837 stone candidates, 149 (17.8%) of them belonging to 99 patients are determined to be false positives. The median number of false positives per scan is 1 (maximum: 8, IQR: [1]). Additionally, 13 stones from 12 patients were missed by the model. Common false positives included artifacts (beam hardening from contrast media, for example), atherosclerotic plaque in the renal sinus or renal artery and calcification or debris related to renal cysts.

3.3 Performance of Stone Tracking

Figure 3 shows three examples of stone tracking, one where the ground truth was inconclusive and another where the predictions were incorrect, and a third where the tracking was correct across 3 time points. Our pipeline achieved a mean MOTA of 0.91 with the 95% confidence interval (0.91, 0.92) on our dataset using our default parameters. Note that in computing MOTA, we only consider patients who have a true stone in at least one scan. In terms of metrics at the stone level, our pipeline achieves edgewise precision and recall rates of 0.82 (0.74, 0.93) and 0.85 (0.81, 0.91), respectively; the precision increased to 0.95 when false positive stones were removed before tracking. In tracking the full trajectory via connected components, we achieve a precision of 0.68 (0.64, 0.73) and recall of 0.88 (0.86, 0.92). The relatively low precision is largely due to false positive singleton stones – removing either all singletons or removing the false positive candidate stones increases the precision to 0.87 (0.83, 0.92), or 0.86 (0.82, 0.91), respectively. In terms of the vertex-based metrics, our method retrieves the correct predecessor with an accuracy of 0.91 (0.88, 0.94) and the correct successor (same stone in the following scan if it exists) with an accuracy of 0.91 (0.88, 0.95).

We also highlight that the performance is quite robust to a wide range of hyperparameter (padding and distance threshold) values, see for example, Fig. 2b – barring the very low range of distance threshold, the performance remains quite stable.

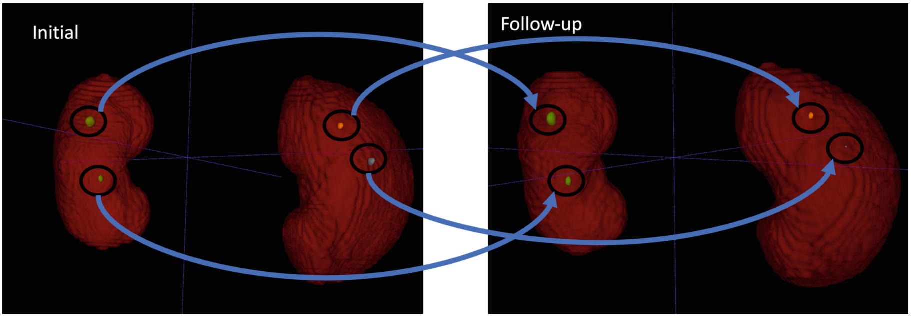

Fig. 4. 3D visualization of matched calculi between the initial and follow-up scans with the follow-up interval of 23 months. The kidney and calculi segmentations as well as the stone matching (indicated by the blue arrows) was performed by our pipeline.

4 Discussion

We have presented a fully automated pipeline to detect and track kidney stones on CTC scans of asymptomatic patients. First, the stones are detected and segmented separately on each scan using a pre-trained deep learning model, and second, the stones found on scans from two different time points are matched to each other using affine registration and greedy matching based on distance in the registered scans Fig. 4. By using affine registration of patches around the kidney, we avoid issues due to relative movement of large organs near the kidney, and unintended deformations due to changes in the stone.

Among existing work, perhaps [18], which introduced the Deep Lesion Tracker to track lesions in the DeepLesion [19] dataset is the closest to ours in spirit. In [18], the tracking problem is cast as a prediction problem: given an object (lesion in their case) in the scan from time point 1, the objective is to predict the location of the same object in the scan from time point 2. The end-to-end method leverages both appearance and anatomical information – the anatomical information helps avoid mistakes when the object looks like other objects in the background. Fortunately, kidney stones are typically quite different from the surrounding kidney tissue. Our approach has several advantages. First, our pipeline is modular; for example, we can easily plug in a different (possibly better) kidney stone detector. Uncoupling the two steps also allows for easier training without any serial training datasets. Second, the pipeline is highly explainable and can easily incorporate human-in-the-loop; the output of the detector can be easily reviewed by the radiologist, and corrected, if necessary. Removing the false positive stones improves the tracking performance significantly (see Fig. 2b).

One of the key contributions in this paper is that we introduce new metrics to evaluate performance of tracking. Unlike the CLEAR MOT metrics, these metrics are tailored to the typical tracking use case in medical imaging – 3D images, only a few “frames” or time points, no requirement for near-real-time processing, and unpredictable relative movement of objects during the long follow-up interval. The vertex-based metrics on retrieving a predecessor or a successor are relevant in the clinic where a physician may want to review the stone in multiple time points to make decisions.

Our study has several limitations. First, the ground truth assessments were performed by a single radiologist – effects of inter-reader variability have not been evaluated. Second, the data came from a single institution, and the number and location of the stones were relatively stable without treatment or interventions; the generalizability of the method has not been assessed. However, given the modularity of our pipeline, we believe our method will be robust to inter-institutional effects.

In conclusion, we have presented an automated pipeline to detect, segment and track kidney stones across multiple time points. We also introduced new graph-based metrics to evaluate tracking performance. These metrics decouple detection, segmentation and tracking errors and are particularly suitable to the medical image domain.

Prospect of Application: We envision deploying the proposed automated stone detection and tracking algorithm to the clinic where it can be used to monitor the progression of kidney stones in both asymptomatic and symptomatic patients. We hope it will reduce the burden on radiologists, reduce inter-reader variability and facilitate more accurate serial assessment of kidney stones.

Table 1. Patient characteristics in the final cohort. The numbers represent Median (IQR). *Missing information for 22 patients

Number of patients with stones (n = 330)	Male (n = 191)	Female (n = 117)
Age at first visit*	57 (52, 63)	57 (52, 60)
Height (cm)*	177.8 (175.2, 182.9)	162.6 (160.0, 167.6)
Weight (kg)*	90.7 (81.7, 101.8)	68.0 (56.7, 81.6)
BMI (kg/m^2)*	28.7 (25.7, 31.7)	25.8 (22.7, 30.3)
Number of scans per patient	2 (2, 2)	
Number of scans with stones	510	
Scan date	2004–2016	
Scan interval (months)	63 (60, 73)	

Acknowledgements. This research was supported by the Intramural Research Program of the National Institutes of Health, Clinical Center, and we utilized the computational resources of the National Institutes of Health high-performance computing Biowulf cluster.

References

1. Smith, R.C., et al.: Acute flank pain: comparison of non-contrast-enhanced CT and intravenous urography. Radiology **194**, 789–794 (1995)
2. Preminger, G.M., Vieweg, J., Leder, R.A., Nelson, R.C.: Urolithiasis: detection and management with unenhanced spiral CT–a urologic perspective. Radiology **207**, 308–309 (1998)
3. Rajapaksa, R.C., Macari, M., Bini, E.J.: Prevalence and impact of extracolonic findings in patients undergoing CT colonography. J. Clin. Gastroenterol. **38**, 767–771 (2004)
4. Hara, A.K., Johnson, C.D., MacCarty, R.L., Welch, T.J.: Incidental extracolonic findings at CT colonography. Radiology **215**, 353–357 (2000)
5. Boyce, C.J., Pickhardt, P.J., Lawrence, E.M., Kim, D.H., Bruce, R.J.: Prevalence of urolithiasis in asymptomatic adults: objective determination using low dose noncontrast computerized tomography. J. Urol. **183**, 1017–1021 (2010)
6. Kang, H.W., et al.: Natural history of asymptomatic renal stones and prediction of stone related events. J. Urol **189**, 1740–1746 (2013)
7. Koh, L.T., Ng, F.C., Ng, K.K.: Outcomes of long-term follow-up of patients with conservative management of asymptomatic renal calculi. BJU Int. **109**, 622–625 (2012)
8. Curhan, G.C.: Epidemiology of stone disease. Urol. Clin. North Am. **34**, 287–293 (2007)
9. Pearle, M.S., Calhoun, E.A., Curhan, G.C.: Urologic diseases of America, P.: urologic diseases in America project: urolithiasis. J. Urol. **173**, 848–857 (2005)
10. Gluecker, T.M., et al.: Extracolonic findings at CT colonography: evaluation of prevalence and cost in a screening population. Gastroenterology **124**, 911–916 (2003)
11. Kampa, R.J., Ghani, K.R., Wahed, S., Patel, U., Anson, K.M.: Size matters: a survey of how urinary-tract stones are measured in the UK. J. Endourol. **19**, 856–860 (2005)
12. Lidén, M., Andersson, T., Geijer, H.: Making renal stones change size—impact of CT image post processing and reader variability. Eur. Radiol. **21**, 2218–2225 (2011)

13. Acosta, J.N., Falcone, G.J., Rajpurkar, P.: The need for medical artificial intelligence that incorporates prior images. Radiology, 212830
14. Elton, D.C., Turkbey, E.B., Pickhardt, P.J., Summers, R.M.: A deep learning system for automated kidney stone detection and volumetric segmentation on noncontrast CT scans. Med. Phy. **49**, 2545–2554 (2022)
15. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear MOT metrics. EURASIP J. Image Video Process. **2008**(1), 1 (2008). https://doi.org/10.1155/2008/246309
16. Pickhardt, P.J., et al.: Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. N Engl J Med **349**, 2191–2200 (2003)
17. Yushkevich, P.A., Pluta, J., Wang, H., Wisse, L.E.M., Das, S., Wolk, D.: IC-P-174: Fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 tesla and 7 tesla T2-weighted MRI. Alzheimers Dement. **12**, P126–P127 (2016)
18. Cai, J., et al.: Deep lesion tracker: monitoring lesions in 4d longitudinal imaging studies. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2021)
19. Yan, K., Wang, X., Lu, L., Summers, R.M.: DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J. Med. Imaging **5**, 1 (2018)