Artificial intelligence-based versus manual assessment of prostate cancer in the prostate gland

a method comparison study

Mortensen, Mike Allan; Borrelli, Pablo; Poulsen, Mads Hvid; Gerke, Oke; Enqvist, Olof; Ulén, Johannes; Trägårdh, Elin; Constantinescu, Caius; Edenbrandt, Lars; Lund, Lars; Høilund-Carlsen, Poul Flemming

Published in:
Clinical Physiology and Functional Imaging

DOI:
10.1111/cpf.12592

Publication date:
2019

Document version:
Accepted manuscript

Citation for pulished version (APA):
Mortensen, M. A., Borrelli, P., Poulsen, M. H., Gerke, O., Enqvist, O., Ulén, J., Trägårdh, E., Constantinescu, C., Edenbrandt, L., Lund, L., & Høilund-Carlsen, P. F. (2019). Artificial intelligence-based versus manual assessment of prostate cancer in the prostate gland: a method comparison study. *Clinical Physiology and Functional Imaging*, *39*(6), 399-406. https://doi.org/10.1111/cpf.12592

DR MIKE ALLAN MORTENSEN (Orcid ID : 0000-0002-7065-9623)
DR ELIN TRÄGÅRDH (Orcid ID : 0000-0002-7116-303X)

Article type : Original Article

Artificial intelligence-based versus manual assessment of prostate cancer in the prostate gland: a method comparison study

Mortensen, Mike Allan [1,2]; Borrelli, Pablo [3]; Poulsen, Mads Hvid [1]; Gerke, Oke [4]; Enqvist, Olof [5]; Ulén, Johannes [6]; Trägårdh, Elin [7,8]; Constantinescu, Caius [4]; Edenbrandt, Lars [3]; Lund, Lars [1,2]; Høilund-Carlsen, Poul Flemming [2,4]

1: Department of Urology, Odense University Hospital, Odense, Denmark

2: Department of Clinical Research, University of Southern Denmark, Odense, Denmark

3: Department of Clinical Physiology, Sahlgrenska University Hospital, Gothenburg, Sweden

4: Department of Nuclear Medicine, Odense University Hospital, Odense, Denmark

5: Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

6: Eigenvision AB, Malmö, Sweden

7: Department of Medical Imaging and Physiology, Skåne University Hospital, Sweden

8: Department of Translational Medicine, Lund University, Malmö, Sweden

Short title: CNN for prostate cancer

Corresponding author

Mike A. Mortensen

Department of Urology

Odense University Hospital

J. B. Winsløwsvej 4, 5000 Odense C, Denmark

E-mail: Mike.allan.mortensen@rsyd.dk

Tel: 0045 40239441

Summary

Aim: To test the feasibility of a fully automated artificial intelligence-based method providing PET measures of prostate cancer (PCa).

Methods: A convolutional neural network (CNN) was trained for automated measurements in ^{18}F-choline (FCH) PET/CT scans obtained prior to radical prostatectomy (RP) in 45 patients with newly diagnosed PCa. Automated values were obtained for prostate volume, maximal standardized uptake value (SUV_{max}), mean standardized uptake value of voxels considered abnormal (SUV_{mean}) and volume of abnormal voxels (Vol_{abn}). The product SUV_{mean} x Vol_{abn} was calculated to reflect total lesion uptake (TLU). Corresponding manual measurements were performed. CNN-estimated data were compared with the weighted surgically removed tissue specimens and manually derived data and related to clinical parameters assuming that 1 g ≈ 1 ml of tissue.

Results: The mean (range) weight of the prostate specimens was 44 g (20-109), while CNN-estimated volume was 62 ml (31-108) with a mean difference of 13.5 g or ml (95% CI: 9.78 – 17.32). The two measures were significantly correlated (r=0.77, p<0.001). Mean differences (95% CI) between CNN-based and manually derived PET measures of SUVmax, SUVmean, Vol_{abn} (ml) and TLU were 0.37 (-0.01 - 0.75), -0.08 (-0.30 – 0.14), 1.40 (-2.26 – 5.06) and 9.61 (-3.95 – 23.17), respectively. PET findings Vol_{abn}, and TLU correlated with PSA (p<0.05), but not with Gleason score or stage.

Conclusion: Automated CNN-segmentation provided in seconds volume and simple PET measures similar to manually derived ones. Further studies on automated CNN-segmentation with newer tracers such as radiolabelled prostate-specific membrane antigen are warranted.

Keywords: Prostatic Neoplasms, -Diagnostic imaging, -Positron emission tomography, -Choline, -Convolutional neural network, -Agreement.

INTRODUCTION

Prostate cancer (PCa) is a heterogeneous disease with an often unpredictable outcome after radical prostatectomy (RP) (Reese et al., 2012). Preoperative risk stratification is often based on prostate specific antigen (PSA) level, Gleason score and clinical stage as suggested by D’Amico (D'Amico et al., 1998). Recurrence-free survival after surgery for clinically localized low risk disease is high. Despite this, around 1 in 4 operated patients will experience biochemical recurrence defined as a rising PSA despite definite therapy (Han et al., 2003). It is well known that PSA in itself is a sub-optimal marker for PCa stage and that determination of both Gleason score, and especially clinical stage, is somewhat subjective. In order to improve appropriate treatment planning and inform patients and relatives better about prognosis, objective and reproducible risk biomarkers are needed. In patients with bone metastases, the use of the automated bone scan index is now a validated prognostic biomarker (Armstrong et al., 2018). With the increasing use of positron emission tomography/computed tomography (PET/CT), methods for automated volumetric calculations of bone metastatic burden are emerging (Lindgren Belal et al., 2017) and direly needed as manual segmentation and determination is both very time consuming and observer-dependent. Imaging is frequently known as a biomarker in patients with disseminated disease, but little is known about this approach in patients undergoing treatment with curative intent. Among multiple PCa tracers, 18F-Choline (FCH) is the one that has been studied most intensively. According to a recent review, the use of PET/CT for prognostication in patients undergoing RP has primarily focused on diagnosing lymph node metastases (Giovacchini et al., 2017), whereas evaluation of the prostate gland itself, correlation with histopathological findings, and impact on outcome have only been touched in relatively few studies (Farsad et al., 2005, Kwee et al., 2006, Kwee et al., 2008, Beheshti et al., 2010). Common for all PET-studies of PCa, is the reliance on visual evaluation susceptible to subjective interpretation. In recent years, however, artificial intelligence (AI) has made its advent into the field of medical imaging (LeCun et al., 2015). By means of AI, it is possible to perform automated pattern recognition and image interpretation within seconds, which is exactly what is required in medical imaging (Litjens et al., 2017). The aim of this study was to compare a fully automated artificial intelligence-based method to manual measurements for measurement of prostatic FCH uptake and to study its correlation with clinical data and post-operative outcome in patients undergoing RP.

METHODS

PATIENTS

Training Group

The training group comprised 145 male patients, who had undergone PET/CT at either Sahlgrenska University Hospital, Gothenburg, Sweden or Odense University Hospital, Odense, Denmark. Ethical approval for the training group was granted by the Regional Ethical Review Boards in Sweden (295-08; 2016/103) and Denmark (3-3013-1692/1).

Study Group

From January 2013 to May 2016, 45 patients underwent FCH-PET/CT prior to RP at Odense University Hospital in Denmark. The included patients had a median age of 67 years (53-75) and pre-operative disease characteristics with a median PSA of 11.0 ng/ml (1.4-43.0) and a median Gleason score of 7 (6-9). Nearly three quarters had palpable tumours in the prostate staged as clinical T2 or T3 (Amin et al., 2017). All patients underwent robot-assisted radical prostatectomy within 3 months of FCH-PET/CT with surgical approach in accordance with current European guidelines (Mottet et al., 2017).

Included patients were part of a larger study of 145 patients undergoing FCH-PET/CT as initial staging in newly diagnosed PCa (Mortensen et al., 2019). Ethical approval for the study group was granted by the Regional Ethical Review Board (S-20120047). The study was approved by the Danish Data Protection Agency and registered at clinicaltrials.gov (NCT02232685). Written informed consent was obtained from all participants.

PET/CT IMAGING

Training group

PET/CT data were acquired using an integrated PET/CT scanner (Siemens Biograph 64 Truepoint or Discovery VCT, GE Healthcare). A low dose CT scan (64-slice helical, 120 kV, 30 mAs, 512x512 matrix) was obtained from the base of the skull to mid-thigh. The CT slice thickness was 5 mm.

Study Group

PET/CT data were acquired using an integrated PET/CT scanner (Discovery VCT, Discovery STE, Discovery RX or Discovery 690, GE Healthcare). A helical diagnostic CT-scan was acquired with in-vivo contrast (ultravist 370 I/ml) using a standard CT protocol (64-slice helical, 120 kV, 'smart mA' maximum 400 mA). Attenuation correction was based on the CT-scan. Patients fasted for 6 hours prior to administration of tracer, each patient receiving a dose of 4 MBq per kg body weight. FCH was produced on automated synthesis systems via alkylation of dimethylaminoethanol with ^{18}F-fluorobromomethane obtaining a radiochemical purity >99%.

IMAGE PROCESSING AND INTERPRETATION

AI-model

In the last few years, convolutional neural networks (CNNs) have revolutionized the field of image analysis, and they are now the first option for image classification, detection and segmentation. CNNs have already been trained to accurately segment organs in CT (Lindgren Belal et al., 2017, Roth et al., 2018). However, due to misalignment between the PET and CT modalities, a direct transfer of a CT-based prostate segmentation to the PET image often leads to voxels in the bladder being incorrectly classified as prostate. As bladder voxels have very high uptake, this could lead to an incorrect analysis. In contrast, this work uses a CNN that simultaneously estimates the misalignment and computes a segmentation that is consistent with both image modalities. We will assume that the misalignment can be approximated as a rigid transformation of the prostate gland, that is, rotation plus translation. The proposed method requires approximate knowledge of the position of the prostate in CT. This is achieved by first running a simple segmentation network on the CT image (Lindgren Belal et al., 2017, Roth et al., 2018). Figure 1 shows the overall structure of our CNN. Inputs to the model are the PET and CT images. The first part of the CNN, the alignment module, takes 50×150×150-subpatches of the PET and CT images roughly centred at the prostate as input and computes a rigid transformation (T) that aligns the prostate in the PET image to the prostate in the CT image. This part of the model is taken from the spatial transformer networks described by Jaderberg et al. (Jaderberg et al., 2015). The resulting transformation is applied to the PET image to produce a CT-aligned PET image. Together with the CT image, this CT-aligned PET image is input to the second part of the CNN, the segmentation module. This is more of a standard segmentation network, with structure as explained in Figure 2. The structure is similar to the popular U-Net but modified to minimize memory usage during training (Ronneberger et al., 2015) . The final convolutional layer has a single output channel with sigmoid activation. For each voxel, the output value describes the estimated probability of that voxel belonging to the prostate. As PET-CT misalignment has been dealt with, this single prostate probability map should be consistent with both modalities. By applying the inverse estimated transformation, T^{-1}, we also get a PET-aligned segmentation. Hence, the outputs from our model are consistent segmentations of both the PET and the CT image (under the assumption that a rigid transformation is sufficient to describe the misalignment).

Training the CNN

The advantage of the proposed architecture is that the network can be trained to focus on the motion of the prostate, while ignoring the possibly inconsistent motion of surrounding organs. To achieve this, we use manual delineations of the prostate made independently in the PET and CT modalities and train the network end-to-end to output segmentations which are consistent with these manual delineations. This means that we can train the alignment module to detect the correct transformation, without knowing it ourselves and it implicitly tells the network to ignore motion in the surrounding tissue.

Since the urinary bladder has very high PET uptake, overlap with the bladder is especially problematic. To avoid this, the urinary bladder was manually delineated in the PET images and we use a negative log-likelihood loss with an auxiliary loss for false positives overlapping with the urinary bladder. More exactly, if P is the set of voxel indices that were labelled as prostate in the PET image, P^c is its complement, B is the set of voxels labelled as bladder and p_i is the network output probability for pixel i, then the loss over the PET image can be written

$$-\sum_{i \in P} \log p_i - \sum_{i \in P^c} \log(1 - p_i) - 10 \sum_{i \in B} \log(1 - p_i)$$

The same loss but without the bladder part is used for the CT image and the two are added to form the total loss.

The annotated data was divided with 80% in a training group and 20% in a validation group used to tune the training. The optimization was performed using the Adam method with Nesterov momentum (Kingma & Ba, 2014). The learning rate was initialized to 0.0001 and reduced when the validation loss reached a plateau. After a few hours of training, the model was evaluated on the training group and false positives where stored in a special group of hard examples that were sampled more frequently (10% of the samples) when the training was restarted. Training and execution were performed using the Tensorflow and Keras frameworks on a high-end desktop computer.

Biomarkers

Standardized uptake value (SUV) on PET-images was automatically calculated in voxels determined to belong to the prostate by the CNN. SUVs above 2.65 were considered abnormal (Reske et al., 2006). Automated measures were obtained for prostate volume, maximal SUV within the prostate (SUV_{max}), mean SUV of voxels considered abnormal (SUV_{mean}) and volume of abnormal voxels in ml (Vol_{abn}). To reflect total lesion uptake (TLU), the product SUV_{mean} x Vol_{abn} was calculated. The non-automated SUV_{max} and SUV_{mean} values were recorded by an experienced nuclear medicine physician (PB) for comparison with automated measurements.

Results of the CNN were both numerical and graphical. Figure 3 illustrates the graphical output of the CNN.

HISTOLOGICAL EXAMINATION

Pre-operative core biopsies of the prostate were processed according to department procedures including description of Gleason score (Epstein et al., 2016). Prostatectomy specimens were all processed according to routine department procedure and in concordance with recommendations from the International Society of Urological Pathology

(Samaratunga et al., 2011). Specimen weight after removal of the seminal vesicles was recorded. Based on estimated tumour extension on whole-mount sectioning of the prostate and weight of the prostate, an estimated tumour burden was calculated assuming that 1 g ≈ 1 ml of tissue.

STATISTICS

An analysis of agreement between manual and automated measurements was done using Bland-Altman plots for SUV_{max}, SUV_{mean}, Vol_{abn} and TLU (Bland & Altman, 1986, Kottner et al., 2011). Association between automated PET measurements and tumour characteristics was examined for using Spearman's rank correlation coefficient. Level of significance was 5%. All analyses were performed using STATA/IC 15.1 (StataCorp, College Station, Texas, USA).

RESULTS

The median weight of the prostate specimen after removal of the seminal vesicles was 44 g (range 20-109). The median automatically calculated prostate gland volume based on automated segmentation of CT-images was 62 ml (range 31-108). We found significant correlation between the automated volume measurement and the manual weight measurement (r=0.77, p<0.01). Automated prostate volume measurements were generally higher than their corresponding manual weight measurements with a mean difference of 13.5 g or ml (95% CI: 9.78 – 17.32). Bland-Altman difference plots for manual weight measurements and automated volume measurements can be seen in Figure 4.

Automated and manually recorded PET measurements are displayed in Table 1. Good agreement between manual and automated measures was seen for SUV_{max} and SUV_{mean} with estimated mean differences around zero as well as limited scatter on Bland-Altman difference plots (Figure 5, upper figures). When comparing automated and manual SUV_{mean}, the automated algorithm tended to overestimate automated SUV_{mean} at low values and underestimate automated SUV_{mean} at high values, compared to manual SUV_{mean}.

For the slightly more complex measurements of Vol_{abn} and TLU, the mean difference between automated and manual measurements were, on average 1.40 and 9.61, respectively (Table 1), and differences were more scattered with increasing mean values on Bland-Altman difference plots (Figure 5 lower figures).

The processing time with the CNN-based method was less than one minute in all cases.

Manually calculated Vol_{abn} significantly correlated with the estimated tumour weight based on the histopathological examination (r=0.32, p=0.04) whereas the automated measurements of Vol_{abn} did not (r=0.15, p=0.32). We did, however, also find that the agreement between automated measurements of Vol_{abn} and the estimated tumour volume did not differ significantly from the agreement between the manually calculated Vol_{abn} and the estimated tumour volume; mean difference 6.9 ml (95% CI: 2.55 – 11.25) and 5.5 ml (95% CI: 2.58 – 8.42), respectively, p=0.45. Both the automated and manually calculated Vol_{abn} overestimated the size of the tumour compared with the estimated tumour volume based on the histopathological examination. When comparing clinical findings with PET-measurements, only pre-operative PSA correlated with automated calculations of Vol_{abn}

(r=0.37, p=0.01) and TLU (r=0.40, p=0.01), but not with other automated PET measures. PSA correlated significantly with manually calculated TLU (r=0.36, p=0.01), but not with other manual PET measures. No correlation between Gleason score or stage and PET measurements (automated or manual) was found.

DISCUSSION

The present study indicates that the CNN used correctly identifies and segments the prostate gland allowing for volumetric determination and uptake measurements comparable to those obtained by manual reading. We found significant correlation between the CT-based estimated volume of the prostate and actual weight of the pathological specimen. The general overestimation of median volume on automated measurements compared to median weight of the specimens could potentially be explained by the fact that the volume estimates represent *in vivo* whereas the weight measures represent bloodless *ex vivo* conditions. The removal of the seminal vesicles in the preparation of the pathology specimen as well as the chemical fixation of the prostate could also affect the results (Lukacs et al., 2014).

Manual and automated SUV_{max} measurements were identical in most cases and, similarly, automated and manually obtained SUV_{mean} values did not differ much, cf. Figure 5. The reason why the automated method tends to over- and underestimate SUV_{mean} at low and high values, respectively, is unclear and needs further exploration. The few cases with non-identical SUV_{max} values could all be explained by uptake falsely detected as inside the prostate – a problem that could potentially be solved with further training of the CNN.

Larger discrepancies between manual and automated analysis were seen, when comparing the more complex measures $\mathrm{Vol}_{\mathrm{abn}}$ and TLU.

When comparing PET-measurements of the prostate to findings after prostatectomy we found a weak, although significant correlation between the manually calculated $\mathrm{Vol}_{\mathrm{abn}}$ and the estimated tumour burden in the prostate specimen. No correlation was found between automated measurements and the estimated tumour burden in the prostate. Common for both the automated method and the manual calculations was a tendency to overestimate the $\mathrm{Vol}_{\mathrm{abn}}$ compared to the actual tumour burden in the prostate. Whether this is caused by the above-mentioned factors, that areas in the prostate defined as pathological are in fact not diseased, or perhaps both, is so far also unclear. The $\mathrm{SUV}_{\mathrm{max}}$ value of 2.65 chosen to define pathological tissue was the one given by Reske and colleagues (Reske et al., 2006). Their study found an AUC of 0.89 in ROC analysis when using an $\mathrm{SUV}_{\mathrm{max}}$ value of 2.65 and ^{11}C-choline as PET-tracer. Similar studies on FCH are not available, although results from a more recent study suggested a need for a higher $\mathrm{SUV}_{\mathrm{max}}$ to distinguish cancer from benign lesions (Schaefferkoetter et al., 2017). The CNN was not able to predict the actual tumour burden of the prostate when compared to histopathological findings after RP. Other alternatives need to be studied in the future where also thresholds relative to tracer uptake in other organs or to $\mathrm{SUV}_{\mathrm{max}}$ could be of interest. $\mathrm{SUV}_{\mathrm{max}}$ as a metric is both resolution dependent and quite susceptible to patient movement, since it only represents one voxel. Furthermore, differences in SUV calculations by different scanners may also have an impact on methods based on fixed SUV thresholds (Adams et al., 2010). Metrics as peak SUV defined as the hottest cubic centimeter might be more robust in those regards. The reason for choosing $\mathrm{SUV}_{\mathrm{max}}$ in the present study was that it is a tried and tested measure, albeit with obvious shortcomings. Partial volume correction has also not been assessed in this study, something that could be done in further prostate PET/CT studies (Alavi et al., 2018).

We found a significant association between pre-operative PSA and several automated measurements, whereas the automated measurements did not significantly correlate with other clinical features. Our study population was rather homogenous primarily comprising less aggressive tumours (predominantly Gleason score 7). With the homogeneous nature of our population and its limited size, subgroup analysis to assess whether FCH would perform differently in more aggressive tumours, could not be performed.

The use of FCH PET/CT in PCa management is well established in cases with suspected recurrence after definitive therapy and in response evaluation in patients with disseminated disease. In recent years, the use of tracers targeting prostate-specific membrane antigen (PSMA) has emerged as a new promising option. Several studies on the use of PSMA targeting tracers have shown superiority over choline-based tracers in a multitude of settings with higher detection rates of low volume and low-grade disease and higher tumour to background ratios (Eapen et al., 2018). Assessment of medical images traditionally relies on the qualitative evaluation by a trained specialist and can be prone to inter-observer variability influenced by education and experience of the given specialist. Using methods based on computer learning, more objective and reproducible measures can be obtained. The AI-based approach used in this study is not in any way applicable to only FCH PET/CT imaging but is a generic tool that with appropriate training can be applied with any given PET tracer, including the impetuous PSMA probes and with any PET/CT scanner for that matter. Also, it is important to notice that the system could with proper training be used to automatically detect and quantify uptake in other regions of interest, e.g. regional lymph nodes, which could be of great interest in PCa staging and in many other malignancies. The method would also be directly applicable in PET/magnetic resonance imaging. As with other methods in machine learning, results may potentially improve with further training. In the model used in the present study, further training might very well result in even better, i.e. more accurate delineation of the prostate gland and cancerous tissue in it, thereby providing a more reliable and clinically more useful tool for the management of PCa.

CONCLUSION

Automated segmentation using an appropriately trained CNN appears to be a feasible and robust tool for automatic segmentation of the prostate gland providing valuable PET measurements in seconds which are similar to more cumbersome manually derived measures. For more accurate and precise measurements, studies applying more highly trained networks are warranted.

The AI-based SUV-measurements as well as those obtained manually were in general not associated with clinical and histopathological findings. The AI-based method can easily be applied with other tracers including PSMA-probes with their higher specificity for PCa lesions.

Acknowledgments

The study was funded by the Danish Cancer Society and the Region of Southern Denmark. The funding bodies had no influence on the design of the study, data collection, data analysis or in writing of the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

REFERENCES

Adams MC, Turkington TG, Wilson JM and Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. *AJR Am J Roentgenol* (2010); **195**: 310-20.

Alavi A, Werner TJ, Hoilund-Carlsen PF and Zaidi H. Correction for Partial Volume Effect Is a Must, Not a Luxury, to Fully Exploit the Potential of Quantitative PET Imaging in Clinical Oncology. *Mol Imaging Biol* (2018); **20**: 1-3.

Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, Gershenwald JE, Compton CC, Hess KR, Sullivan DC, Jessup JM, Brierley JD, Gaspar LE, Schilsky RL, Balch CM, Winchester DP, Asare EA, Madera M, Gress DM and Meyer LR (2017). AJCC Cancer Staging Manual. Cham, Switzerland, Springer International Publishing.

Armstrong AJ, Anand A, Edenbrandt L, Bondesson E, Bjartell A, Widmark A, Sternberg CN, Pili R, Tuvesson H, Nordle O, Carducci MA and Morris MJ. Phase 3 Assessment of the Automated Bone Scan Index as a Prognostic Imaging Biomarker of Overall Survival in Men With Metastatic Castration-Resistant Prostate Cancer: A Secondary Analysis of a Randomized Clinical Trial. *JAMA Oncol* (2018); **4**: 944-51.

Beheshti M, Imamovic L, Broinger G, Vali R, Waldenberger P, Stoiber F, Nader M, Gruy B, Janetschek G and Langsteger W. 18F choline PET/CT in the preoperative staging of prostate cancer in patients with intermediate or high risk of extracapsular disease: a prospective study of 130 patients. *Radiology* (2010); **254**: 925-33.

Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* (1986); **1**: 307-10.

D'Amico AV, Whittington R, Malkowicz SB, Schultz D, Blank K, Broderick GA, Tomaszewski JE, Renshaw AA, Kaplan I, Beard CJ and Wein A. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *Jama* (1998); **280**: 969-74.

Eapen RS, Nzenza TC, Murphy DG, Hofman MS, Cooperberg M and Lawrentschuk N. PSMA PET applications in the prostate cancer journey: from diagnosis to theranostics. *World J Urol* (2018).

Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR and Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. *Am J Surg Pathol* (2016); **40**: 244-52.

Farsad M, Schiavina R, Castellucci P, Nanni C, Corti B, Martorana G, Canini R, Grigioni W, Boschi S, Marengo M, Pettinato C, Salizzoni E, Monetti N, Franchi R and Fanti S. Detection and localization of prostate cancer: correlation of (11)C-choline PET/CT with histopathologic step-section analysis. *J Nucl Med* (2005); **46**: 1642-9.

Giovacchini G, Giovannini E, Leoncini R, Riondato M and Ciarmiello A. PET and PET/CT with radiolabeled choline in prostate cancer: a critical reappraisal of 20 years of clinical studies. *Eur J Nucl Med Mol Imaging* (2017); **44**: 1751-76.

Han M, Partin AW, Zahurak M, Piantadosi S, Epstein JI and Walsh PC. Biochemical (prostate specific antigen) recurrence probability following radical prostatectomy for clinically localized prostate cancer. *J Urol* (2003); **169**: 517-23.

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. Spatial Transformer Networks. *eprint arXiv:1506.02025* (2015): arXiv:1506.02025.

Kingma DP and Ba J. Adam: A Method for Stochastic Optimization. *eprint arXiv:1412.6980* (2014): arXiv:1412.6980.

Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, Roberts C, Shoukri M and Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* (2011); **64**: 96-106.

Kwee SA, Thibault GP, Stack RS, Coel MN, Furusato B and Sesterhenn IA. Use of step-section histopathology to evaluate 18F-fluorocholine PET sextant localization of prostate cancer. *Mol Imaging* (2008); **7**: 12-20.

Kwee SA, Wei H, Sesterhenn I, Yun D and Coel MN. Localization of primary prostate cancer with dual-phase 18F-fluorocholine PET. *J Nucl Med* (2006); **47**: 262-9.

LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature* (2015); **521**: 436-44.

Lindgren Belal S, Sadik M, Kaboteh R, Hasani N, Enqvist O, Svarm L, Kahl F, Simonsen J, Poulsen MH, Ohlsson M, Hoilund-Carlsen PF, Edenbrandt L and Tragardh E. 3D skeletal uptake of (18)F sodium fluoride in PET/CT images is associated with overall survival in patients with prostate cancer. *EJNMMI Res* (2017); **7**: 15.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak J, van Ginneken B and Sanchez CI. A survey on deep learning in medical image analysis. *Med Image Anal* (2017); **42**: 60-88.

Lukacs S, Vale J and Mazaris E. Difference between actual vs. pathology prostate weight in TURP and radical robotic-assisted prostatectomy specimen. *International Brazilian JOurnal of Urology* (2014); **40**: 823-7.

Mortensen MA, Poulsen H, Gerke O, Jakobsen JS, Høilund-Carlsen PF and Lund L. 18F-fluoromethylcholine-PET/CT for diagnosing bone and lymph node metastases in patients with intermediate- or high-risk prostate cancer. *Prostate International* (2019); **In press**.

Mottet N, Bellmunt J, Bolla M, Briers E, Cumberbatch MG, De Santis M, Fossati N, Gross T, Henry AM, Joniau S, Lam TB, Mason MD, Matveev VB, Moldovan PC, van den Bergh RCN, Van den Broeck T, van der Poel HG, van der Kwast TH, Rouviere O, Schoots IG, Wiegel T and Cornford P. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol* (2017); **71**: 618-29.

Reese AC, Pierorazio PM, Han M and Partin AW. Contemporary evaluation of the National Comprehensive Cancer Network prostate cancer risk classification system. *Urology* (2012); **80**: 1075-9.

Reske SN, Blumstein NM, Neumaier B, Gottfried HW, Finsterbusch F, Kocot D, Moller P, Glatting G and Perner S. Imaging prostate cancer with 11C-choline PET/CT. *J Nucl Med* (2006); **47**: 1249-54.

Ronneberger O, Fischer P and Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *eprint arXiv:1505.04597* (2015): arXiv:1505.04597.

Roth HR, Shen C, Oda H, Oda M, Hayashi Y, Misawa K and Mori K. Deep Learning and Its Application to Medical Image Segmentation. *Medical Imaging Technology* (2018); **36**: 63-71.

Samaratunga H, Montironi R, True L, Epstein JI, Griffiths DF, Humphrey PA, van der Kwast T, Wheeler TM, Srigley JR, Delahunt B and Egevad L. International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 1: specimen handling. *Mod Pathol* (2011); **24**: 6-15.

Schaefferkoetter JD, Wang Z, Stephenson MC, Roy S, Conti M, Eriksson L, Townsend DW, Thamboo T and Chiong E. Quantitative (18)F-fluorocholine positron emission tomography for prostate cancer: correlation between kinetic parameters and Gleason scoring. *EJNMMI Res* (2017); **7**: 25.

Table 1 Comparison of automated and manual measurements

Variable	Automated,	Manual,	Mean difference	BA LoA
	median (range)	median (range)	(95% CI)	
SUV_{max}	8.0 (2.7-15.5)	7.5 (2.7-15.5)	0.37 (-0.01 - 0.75)	-2.83 - 2.09
SUV_{mean}	3.6 (2.7-5.5)	3.3 (2.0-7.7)	-0.08 (-0.30 – 0.14)	-1.37 - 1.53
Vol_{abn} (ml)	14.4 (0.1-79.6)	11.1 (1.1-49.3)	1.40 (-2.26 – 5.06)	-25.25 - 22.44
TLU	50.6 (0.3-414.2)	40.5 (3.2-202.6)	9.61 (-3.95 – 23.17)	-98.07 - 78.84

BA LoA: Bland-Altman Limits of Agreement

Figure Legends

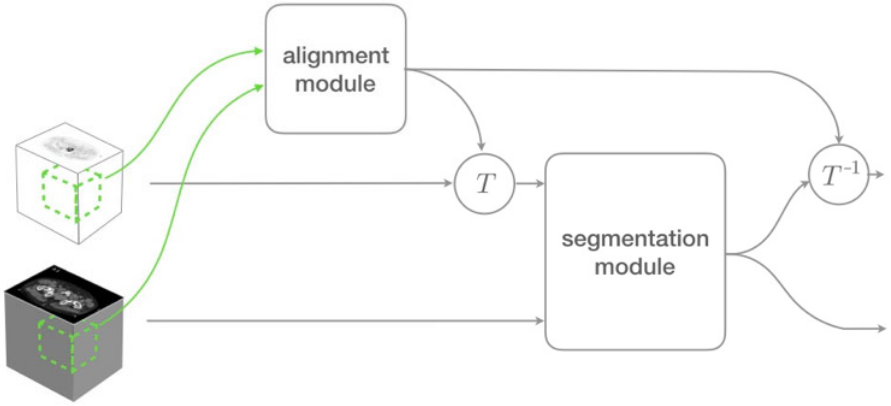
Figure 1: Model structure. The first part of the model obtains sub-patches of the images roughly centered at the prostate and uses these to compute a transformation aligning the PET image to the CT image. This allows the segmentation module to consider both modalities in the segmentation process.

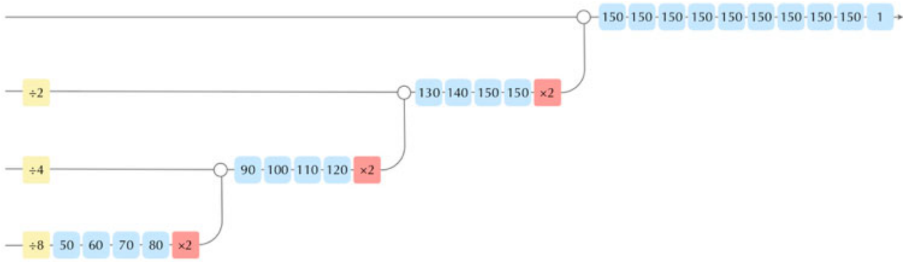
Figure 2: Segmentation module. Blue boxes are 3×3×3 convolutional layers and the number indicates the number of filters. Red boxes are 2×-up-sampling layers and yellow boxes are average pooling where the number indicates the pool size. The pooling layers mean that the segmentation module works on four different resolutions. This allows a large receptive field at low memory cost during training. All convolutional layers use rectified linear unit activations apart from the last one using a sigmoid activation to produce the final output probabilities.

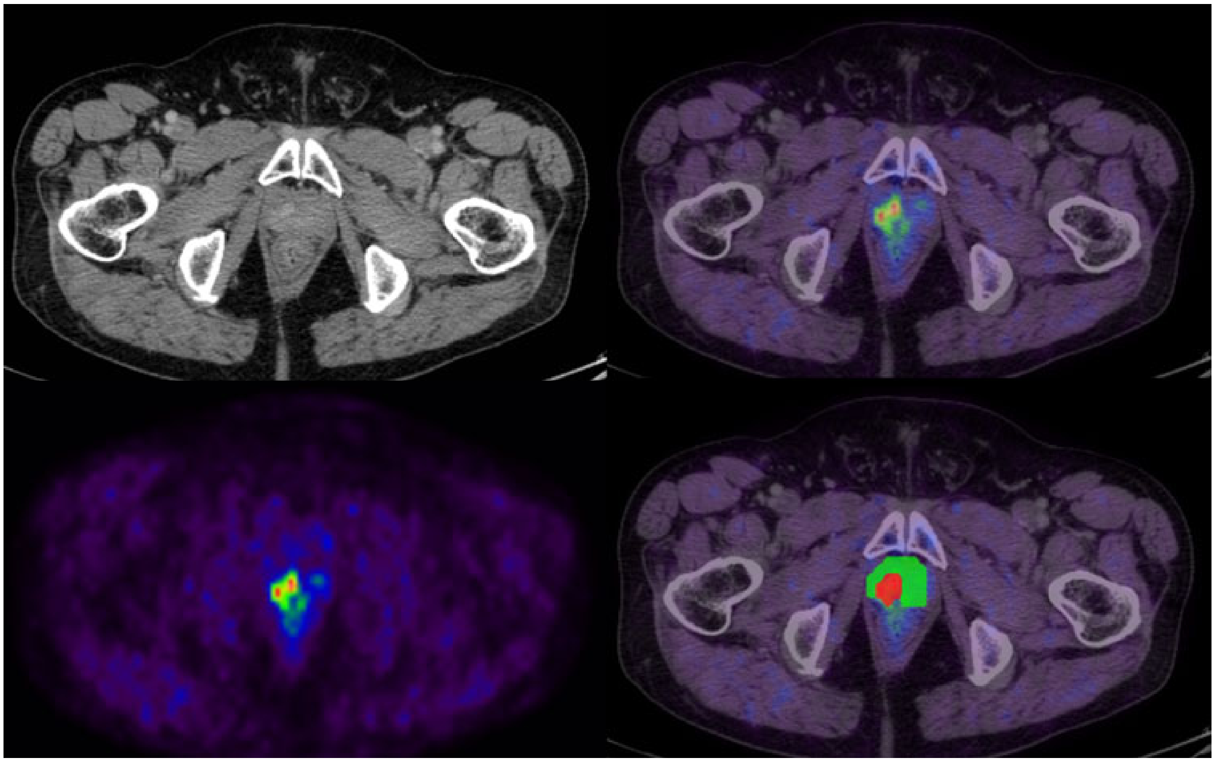
Figure 3: Graphical output of the CNN showing CT-sequence (upper left figure), PET-sequence (lower left figure), PET/CT-fusion (upper right figure) and CNN output (lower right figure) with the segmented prostate (green) and tumour (red).

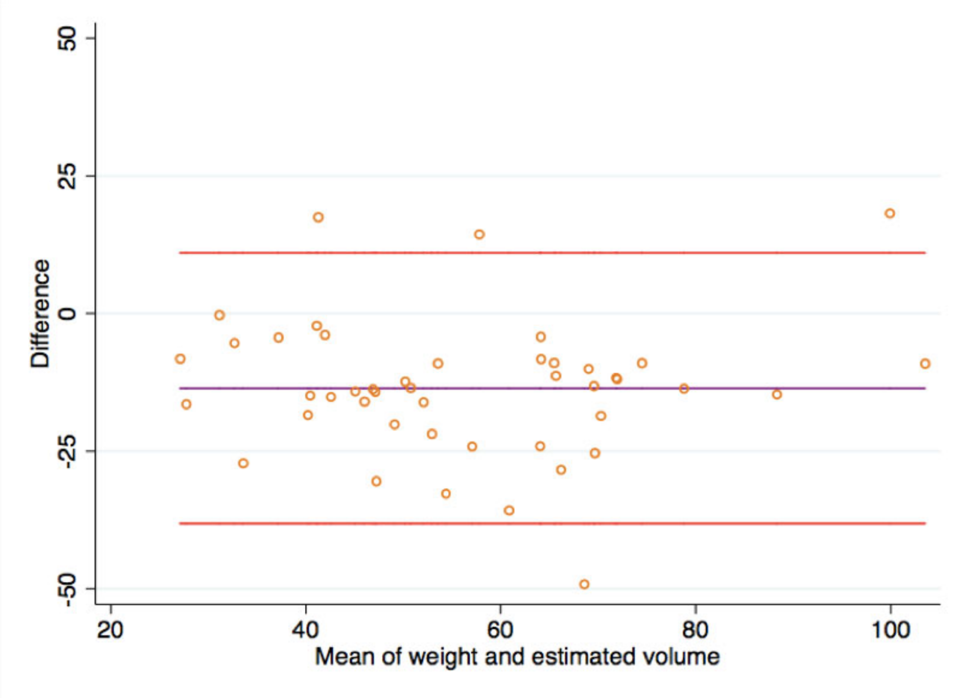
Figure 4: Bland-Altman difference plots of weight and automated prostate volume estimate plotted against the mean of the two methods. The purple line indicates the mean of the differences whereas the red lines indicate the upper and lower limits of agreement (BA LoA).

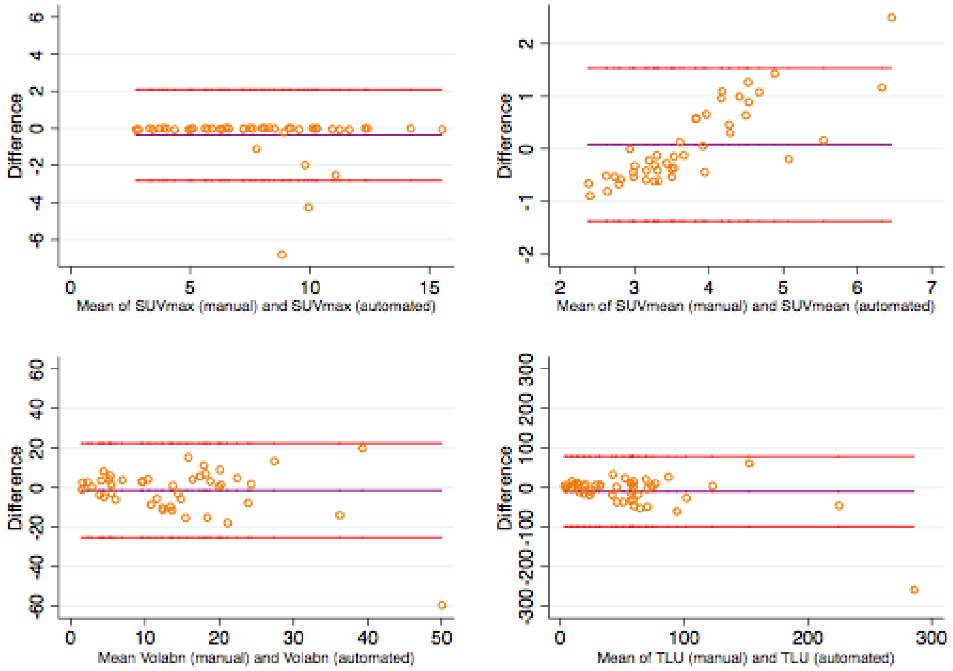
Figure 5: Bland-Altman difference plots of SUV_{max}, SUV_{mean}, Vol_{abn} and TLU plotted against the mean of the two methods. The purple line indicates the mean of the differences whereas the red lines indicate the upper and lower limits of agreement (BA LoA).


alignment module
T
segmentation module
T^{-1}


150
150
150
150
150
150
150
150
150
1
÷2
130
140
150
150
×2
÷4
90
100
110
120
×2
÷8
50
60
70
80
×2




50
25
0
-25
-50
Difference
20
40
60
80
100
Mean of weight and estimated volume


Difference
6
4
2
0
-2
-4
-6
0
5
10
15
Mean of SUVmax (manual) and SUVmax (automated)
Difference
2
1
0
-1
-2
2
3
4
5
6
7
Mean of SUVmean (manual) and SUVmean (automated)
Difference
60
40
20
0
-20
-40
-60
0
10
20
30
40
50
Mean Volabn (manual) and Volabn (automated)
Difference
300
200
100
0
-100
-200
-300
0
100
200
300
Mean of TLU (manual) and TLU (automated)