European Journal of Radiology 91 (2017) 41–46

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

In-depth analysis of interreader agreement and accuracy in categorical assessment of brown adipose tissue in (18)FDG-PET/CT

Anton S. Becker[a,c,*], Caroline Zellweger[b], Khoschy Schawkat[a], Sanja Bogdanovic[a], Valerie Doan Phi van[a], Hannes W. Nagel[b], Christian Wolfrum[c], Irene A. Burger[b]

[a] Institute of Diagnostic and Interventional Radiology, University Hospital of Zurich, Switzerland
[b] Department of Nuclear Medicine, University Hospital of Zurich, Switzerland
[c] Department of Health Science and Technology, ETH Zurich, Switzerland

ARTICLE INFO

Keywords:
Hybrid imaging
PET/CT
Brown fat
Interreader agreement

ABSTRACT

Purpose: To evaluate the interreader agreement of a three-tier craniocaudal grading system for brown fat activation and investigate the accuracy of the distinction between the three grades.

Materials and methods: After IRB approval, 340 cases were retrospectively selected from patients undergoing (18)FDG-PET/CT between 2007 and 2015 at our institution, with 85 cases in each grade and 85 controls with no active brown fat. Three readers evaluated all cases independently. Furthermore standardized uptake values (SUV) measurements were performed by two readers in a subset of 53 cases. Agreement between the readers was assessed with Cohen's Kappa (k), the concordance correlation coefficient (CCC) and the intraclass correlation coefficient (ICC). Accuracy was assessed with Bland-Altman and receiver operating characteristics (ROC) analysis. A Bonferroni-corrected two-tailed $p < 0.016$ was considered statistically significant.

Results: Agreement for BAT grade was excellent by all three metrics with k = 0.83–0.89, CCC = 0.83–0.89 and ICC = 0.91–0.94. Bland-Altman analysis revealed only slight average over- or underestimation (−0.01 − 0.14) with the majority of disagreements within one grade. ROC analysis yielded slightly less accurate classification between higher vs. lower grades (Area under the ROC curves 0.78–0.84 vs. 0.88–0.92) but no significant differences between readers. Agreement was also excellent for the maximum SUV and the total brown fat volume (k = 0.90 and 0.94, CCC = 0.93 and 0.99, ICC = 0.96 and 0.99), but Bland-Altman plots revealed a tendency to underestimate activity by one of the readers.

Conclusion: Grading the activation of brown fat by assessment of the most caudally activated depots results in excellent interreader agreement, comparable to SUV measurements.

1. Introduction

Obesity has become a serious health problem worldwide. Despite all efforts, obesity is still on the rise globally and the prevalence will surpass 18% in men and 21% in women after 2025 [1]. Chronic obesity entails several co-morbidities such as type II diabetes, atherosclerotic changes, early osteoarthritis and some types of cancer [2]. Although various approaches have been undertaken to fight obesity, for example, but not limited to, lifestyle modification or pharmacologic therapy these attempts have, at best, yielded modest results [3,4]. For this reason, the initial discovery of metabolically active brown adipose tissue ('brown fat') in adult humans in 2002 [5] and the subsequent description in a larger human cohort seven years later [6,7] has sparked an active field of research on how to exploit its “calorie burning properties” to facilitate weight loss. The main function of brown fat, as opposed to white fat, which stores excess energy in the form of triacylglycerols, is the dissipation of energy into heat. For this purpose triacylglycerols, amino acids or glucose can be utilized. Thanks to the latter substrate, brown fat can be detected in-vivo by (18)FDG-PET/CT due to its ability to precisely match functional metabolic with spacial anatomic information. To date, (18)FDG-PET hybrid imaging with CT or MRI is the only medical imaging modality able to do so. Beyond detection, (18)FDG-PET also allows quantification of glycolytic activity in standarized uptake values (SUV). This is of paramount importance in order to be able to evaluate new therapeutic approaches. Bahler and colleagues have recently demonstrated excellent interreader agreement of brown fat SUV measurements [8]. However, the measurement process of brown fat tends to be cumbersome and time-consuming because other glycolytic tissues (myocardium, liver, muscle) need to be manually cropped from the volume of interest. Albeit there are

* Corresponding author at: Institute of Diagnostic and Interventional Radiology, University Hospital of Zurich, Raemistrasse 100, Switzerland.
E-mail address: anton.becker@usz.ch (A.S. Becker).

http://dx.doi.org/10.1016/j.ejrad.2017.03.012
Received 5 December 2016; Received in revised form 14 March 2017; Accepted 21 March 2017
0720-048X/

interesting approaches to automate this process such as the one presented by Gifford et al. [9], they are usually fairly complicated per-se and not readily available. Moreover, specialized software is needed which is usually associated with high costs and a steep learning curve. Recently, it was shown that more caudal activation of brown fat correlates with all glycolytic measures: SUV_{max}, metabolic active volume as well as total brown fat activity [10], which led to the proposal of a three-tier grading system based on the most caudally activated brown fat depot. Such a system allows for faster classification on a standard radiology workstation. Especially for the analysis of larger study cohorts, which would be necessary to demonstrate the efficacy of brown fat activation in the treatment of obesity, this could substantially simplify read outs. However, it is not clear what impact such a grading system will have on the interreader agreement. The binning of continuous SUV measurements, which are very reliable [8], into discrete categories could theoretically lead to more pronounced disagreement. Thus, the purpose of this study was to evaluate the interreader agreement of the three-tier craniocaudal grading system, compare it to the agreement of SUV measurements and investigate the accuracy of the readers for the distinction between the three grades.

2. Materials and methods

This retrospective study was approved by the IRB, who waived the need for informed consent.

2.1. FDG-PET/CT protocol

All scans had been acquired as routine clinical examinations. In summary, patients were instructed to fast for at least 6 h prior to the examination. Fasting blood glucose was verified to be below < 7 mmol/L. After injection of 4 mBq FDG per kilogram bodyweight, patients were laid with closed eyes in supine position in a quiet, even-tempered room. After 60 min (± 5 min) a low-dose CT scan was acquired (120–140 kV, mAs dynamically regulated by SmartmA (R)), followed by the PET scan from mid-thigh to the vertex of the skull.

2.2. Grading system

The activation strength of the brown fat was divided into four categories according to the grades proposed by [10]: 0 = no active brown fat, 1 = nuchal and/or supraclavicular, 2 = thoracic, 3 = infra-diaphragmal fat depots activated, as illustrated in Fig. 1.

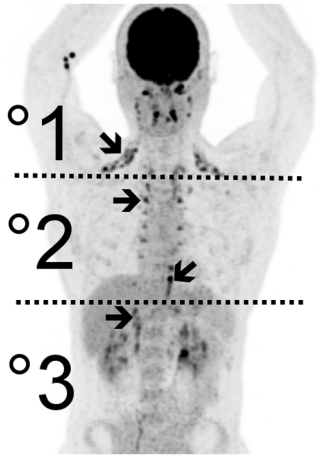

Fig. 1. Maximum intensity projection PET image depicting a grade 3 activation with lines demarcating the different discriminative depots as proposed in [8]. Brown fat exhibits a cranio-caudal activation pattern. In other words, the more caudally the brown fat is activated, the higher the overall and maximum glucose uptake.

2.3. Patient population

In a retrospective analysis of roughly 8300 patients undergoing FDG-PET/CT in the years 2007–2015, metabolically active brown fat was detected in 758 patients. Metabolically active brown fat was defined as fat density tissue on CT (− 250 to − 50 Hounsfield Units) with a corresponding SUV_{max} of ≥ 2.5 g/ml. All cases were classified according to the abovementioned three-tier anatomical grading system. From each category 85 cases were randomly selected, resulting in 340 cases. From the electronic patient file the indication for the examination was retrieved and coded by tumor type after the ICD-10 system.

2.4. Readout

One reader who had performed the initial screening and readout of all PET/CT examinations (ASB) served as a reference standard. Three readers who were blinded to the patient selection independently performed the readout on a standard radiological worksataion (IMPAX, AGFA HealthCare Inc., Bonn, Germany): KS (4^{th} year radiology resident, no prior experience in PET/CT), SB (1^{st} year radiology resident, completed a four-month rotation in PET/CT) and CZ (board certified radiologist, 2 years experience in PET/CT). Prior to the readout four cases not included in the study population were evaluated as training cases. An instruction sheet with a schematic similar to Fig. 1 was available during the readout.

In a subset of examinations with active brown fat (n = 53), two readers independently performed SUV measurements (HWN and VDPV) using a clinical workstation for hybrid imaging (GE AWServer v.4.7; GE Healthcare, Chigaco, IL, USA). The following measurements were performed: maximum SUV (SUV_{max}), mean SUV (SUV_{mean}) and metabolically active brown fat volume (MBFV). Activity below 2.5 g/ml was defined as background activity (fixed threshold for MBFV and SUV_{mean}).

2.5. Statistical analysis

Statistical Analysis was performed in R version 3.3.2 (R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org/). Graphs were produced with ggplot2 [11]. The full data and analysis are available in an online repository at http://github.com/ASBecker/BAT-IRR.

2.5.1. Interreader agreement

Interreader agreement for all reader-pairs was investigated with Cohen's weighted kappa (κ), which accounts for agreements by chance [12], and with the concordance correlation coefficient (CCC) [13] and the intraclass correlation coefficient (ICC) 2k [14]. Since all three metrics are widely used in the radiological literature we opted to calculate and report each to allow better comparability to other studies, which commonly only use one. Scores were considered significantly different if the 95% confidence intervals (CIs) did not overlap and valued as follows: slight (< 0.20), fair (0.20–0.39), moderate (0.40–0.59), substantial (0.60–0.79), and excellent (> 0.80) agreement.

2.5.2. Classification accuracy

The deviation from the reference classification was assessed with Bland-Altman analysis. Classification accuracy between the subsequent categories (0–1, 1–2 and 2–3) was assessed with a receiving operator characteristics (ROC) analysis. The area under the ROC-curve (A_z) was compared with DeLong's nonparametric test [15]. A p-value of < 0.016 (< 0.05 with Bonferroni correction for 3 combinations, two-tailed alpha) was considered statistically significant.