Original article

Nuclear Medicine Communications

Observer agreement of treatment responses on planar bone scintigraphy in prostate cancer patients: importance of the lesion assessment method

Randi F. Fonager[a], Helle D. Zacho[a,c], Signe Albertsen[b], Joan Fledelius[e], June A. Ejlersen[e], Mette H. Christensen[f], Ramune Aleksyniene[a], José A. Biurrun Manresa[d] and Lars J. Petersen[a,c]

Purpose **The aim of this study was to assess observer agreement on the evaluation of treatment responses of bone metastases by bone scintigraphy (BS) using different scoring methods in prostate cancer patients.**

Patients and methods **Sixty-three paired BS from 55 patients were included. BS was performed before and after more than 12 weeks of anticancer treatment. A panel of experienced nuclear medicine physicians from several institutions evaluated treatment response using three different methods: (a) standard clinical assessment, (b) MD Anderson criteria, and (c) Prostate Cancer Working Group 2 (PCWG-2) criteria. All methods were based on the evaluation of paired before–after bone scans.**

Results **Readers were able to classify the presence of bone metastases at baseline with a high level of agreement [Cohen's $\kappa = 0.94$, 95% confidence interval (CI) 0.82–1.00]. Observer agreement on bone response by PCWG-2 criteria showed considerable agreement (Cohen's $\kappa = 0.84$, 95% CI: 0.69–0.99). Evaluation using standard clinical assessment and MD Anderson criteria showed moderate agreement (0.52, 95% CI: 0.36–0.69 and 0.64, 95% CI: 0.48–0.79, respectively). There was considerable variation among readers for regional lesion count on individual scans, with limits of agreement of −10 to 10 lesions or more for the majority of anatomical regions, including the thorax, spine, and pelvis.**

Conclusion **Observer agreement on treatment response by BS varied notably across methods. Optimal agreement was achieved by the PCWG-2 criteria. Variation in the classification of treatment response of bone metastases may have a significant impact on clinical decision-making, emphasizing the need for a uniform approach, including during clinical practice. Response assessment by lesion counting on repeated BS without access to previous scans cannot be recommended.** ***Nucl Med Commun*** **38:215–221**

Nuclear Medicine Communications 2017, **38**:215–221

Keywords: bone metastases, bone scan, observer agreement, prostate cancer, response assessment

[a]Department of Nuclear Medicine, Clinical Cancer Research Center, [b]Department of Urology, Aalborg University Hospital, Departments of [c]Clinical Medicine, [d]Health Science and Technology, Aalborg University, Aalborg, [e]Department of Nuclear Medicine, Herning Hospital, Herning and [f]Department of Clinical Physiology, Viborg Regional Hospital, Viborg, Denmark

Correspondence to Randi F. Fonager, Cand. Scient. Med., Department of Nuclear Medicine Aalborg University Hospital, Hobrovej 18-22, DK-9000 Aalborg, Denmark
Tel: +45 9766 5493; fax: +45 9766 5501; e-mail: r.fuglsang@rn.dk

Received 22 July 2016 **Accepted** 16 December 2016

Introduction

Prostate cancer is the second leading cause of cancer deaths in men, and bone metastasis is one of the most frequent and severe causes of morbidity [1,2]. New and upcoming therapies are costly and economically challenging for the healthcare system worldwide. The average costs of treatment with new cancer therapies have more than doubled in recent years [3–5]. As a consequence, it is important that methods for the evaluation of treatment response and patient benefit are valid and precise to stop ineffective and costly treatments in a timely manner.

According to international guidelines, planar whole-body bone scintigraphy (BS) remains the recommended method for staging and evaluating bone metastases [6,7]. Standard clinical assessment often classifies bone changes as regression, stable disease, or progression. In the past, standardized response criteria have focused on providing details for the classification of response categories, for example, the Response Evaluation Criteria in Solid Tumours [8]. However, Response Evaluation Criteria in Solid Tumours defines bone metastases on BS as unmeasurable lesions. In 2004, Hamaoka and colleagues proposed the MD Anderson response criteria specifically for bone metastases. These criteria acknowledge the presence of a response rather than quantifying possible changes [9]. In 2008, the Prostate Cancer Working Group 2 (PCWG-2) proposed criteria to determine the presence of radiographic progression. PCWG-2 defines progression as the appearance of two or more confirmed

0143-3636 DOI: 10.1097/MNM.0000000000000643

bone metastases [10]. The PCWG-2 criteria have since been widely adopted, primarily in clinical trials. However, the consistency of these response evaluation methods has not been investigated systematically.

The knowledge of observer agreement in treatment response assessment on BS is scarce. Kaboteh *et al.* [11] showed an apparently high level of agreement among three experienced readers using the PCWG-2 criteria in 266 patients for the assessment of BS changes. However, the prevalence of bone metastasis was not reported and patients with extensive metastatic disease were excluded. Anand *et al.* [12] included 173 patients with metastatic disease from the study by Kaboteh and colleagues and reported agreement among readers using 'increased burden' and PCWG-2 criteria. They found no difference in observer agreement between the two reporting methods, but large pair-wise observer variations for the assessment of disease progression within methods, with κ values ranging from 0.55 to 0.90. Observer agreement for standard clinical assessment and MD Anderson criteria has not been investigated. Furthermore, no direct comparison of observer agreement for different response assessment methods has been attempted.

The aim of the present study was to assess observer agreement among a panel of experienced nuclear medicine physicians from several institutions evaluating bone metastasis responses on BS in prostate cancer patients undergoing various palliative cancer therapies. We investigated response assessment of paired (before and after therapy) BS using standard clinical assessment, MD Anderson, and PCWG-2 criteria. In addition, we assessed observer agreement for counting regional bone lesions in individual bone scans without comparison with other BS results.

Patients and methods

Patients

We identified all prostate cancer patients who had two or more BS performed within 1 year during the period from 1 January 2009 to 22 November 2014 at the Department of Nuclear Medicine at Aalborg University Hospital, Denmark. The following eligibility criteria were applied: (a) The patients were treated with androgen-deprivation therapy (ADT), next-generation hormonal therapy (NGH) (abiraterone or enzalutamide), or chemotherapy and (b) each patient had two successive BS, a baseline scan, and a follow-up scan while on the same treatment. The baseline BS had to be performed within 3 months before therapy and a maximum of 14 days after the initiation of treatment. The follow-up BS had to be performed from 12 to 52 weeks within treatment for ADT or 12–30 weeks within treatment for NGH and chemotherapy. Patients treated successively by different treatments regimens and having more than one BS pair fulfilling the eligibility criteria were allowed to be included twice.

Bone scintigraphy

Bone scans were performed in accordance with the European Society of Nuclear Medicine guideline on BS [13]. Acquisition of whole-body, planar BS was performed 2 h after an intravenous injection of 750–1000 MBq ^{99}Tc-labeled hydroxymethylene diphosphonate using a dual-head gamma camera with simultaneous anterior and posterior whole-body acquisition and a multipurpose low-energy high-resolution collimator according to institutional practices. Any additional single-photon emission tomography/computed tomography was not included in the response assessment.

Bone scintigraphy reading and classification

Five board-certified specialists in nuclear medicine with 4–13 years of experience participated in reading the BS. Readers were recruited from three institutions to minimize within-institution reading habituation. The readers were blinded from the original interpretation of the BS and any clinical data, except date of birth and the prostate cancer diagnosis.

Baseline and follow-up images were evaluated independently by the same two readers and assessed on a dichotomous scale: bone metastases present (M1) or not (M0). In M1 patients, readers were asked to indicate whether the BS was consistent with a super scan.

Establishing progression according to the PCWG-2 criteria depends on the ability of readers to identify new bone metastases and thus counting/identification of individual lesions. Although a side-by-side assessment of images is likely recommended, it is not explicitly specified by PCWG [10]. Therefore, we evaluated observer agreement in lesion counting when looking at baseline and follow-up scans separately. Individual lesions were counted in five skeletal regions: the skull (including face), thorax (ribs and sternum), spine, pelvis (including sacrum), and the extremities (including the shoulder blade). Up to 20 lesions were counted per region. Subsequently, on the basis of the regional lesion counts, images were categorized according to the extent of disease (EOD) classification by Soloway *et al.* [14]: 0 = no bone metastases, 1 = < 6 bone metastases, 2 = 6–20 bone metastases, 3 = > 20 bone metastases, or 4 = super scan.

Bone scan response

Each pair of baseline and follow-up BS were then evaluated side by side for changes during treatment using three different methods (Table 1). Each BS pair was evaluated by two readers; however, for logistical reasons, the same two readers did not assess each bone scan pair by all three response assessment methods. For example, in patient 1, reader 1 and reader 2 assessed baseline and follow-up BS by standard clinical assessment, reader 3 and reader 4 assessed baseline and follow-up BS by MD Anderson criteria, and finally, reader 5 and reader 1 assessed baseline and follow-up BS by PCWG-2 criteria.

Table 1 Treatment response assessment methods

	Standard clinical assessment	MD Anderson	PCWG-2
Nonprogressive disease	–	Complete response Complete disappearance of hot spots or tumor signal	–
	Regression	Partial response Regression of lesions	–
	Stable disease	Stable disease No new lesions	Nonprogressive disease No, or maximum one new lesion
Progressive disease	Progressive disease	Progressive disease New lesions, increased intensity of existing lesions, or both	Progressive disease Two or more new lesions
	Mixed response Some areas showing regression and some areas showing progression	–	–

PCWG-2, Prostate Cancer Working Group 2.

A consensus reading was performed if the two readers disagreed. For each response assessment method, BS pairs were evaluated by five pairs of observers.

Readers received detailed written and oral instructions on the criteria for each of the response assessment methods. For standard clinical assessment, the instruction was to rate the response as they would in a standard clinical situation at their institution and place their rating in one of three response categories (Table 1). In some cases, readers checked both regression and progression, thereby creating a fourth category, hereafter named 'mixed response'. The readers argued that this was often the case in clinical situations. Subsequently, when recategorizing responses as progressive disease (PD) or non-PD, the 'mixed response' category was redefined as PD on the basis of a consensus that if progression was present, even with the presence of regression in other parts of the skeleton, the patient had PD. The MD Anderson criteria were provided with explanatory comments for each of the response categories on the basis of the original description by Hamaoka *et al.* [9] (Table 1) and readers were asked to assess the BS pairs accordingly. The PCWG-2 criteria were likewise provided, and readers were asked to assess the BS pairs accordingly. However, PCWG-2 criteria require that the presence of two or more new lesions must be confirmed on a subsequent scan a minimum of 6 weeks later, and at the time of the study, this was not routine practice at our department; consequently, confirmatory images were not available.

Statistical analysis

Cohen's κ was used to assess observer agreement and reported with 95% confidence intervals (CIs). We used unweighted Cohen's κ for dichotomous and nonordinal scales [15]. For the cases in which the scale was nondichotomous with a clearly ordinal ranking (EOD and MD Anderson response criteria), linear-weighted Cohen's κ was also calculated. The extent of agreement of κ values was interpreted according to the terminology by Landis and Koch [16]: $\kappa = 0.00$–0.20: slight, 0.21–0.40: fair, 0.41–0.60: moderate, 0.61–0.80: substantial, and 0.81–1.00: almost perfect agreement. Agreement on the number of lesions in the different skeletal regions was assessed by Bland–Altman analysis [17]. Bland–Altman plots present the bias, that is, the mean difference in lesion count between readers, and 95% limits of agreement, which provide reference values of the maximum differences that can be expected between readers when counting the same BS.

Approvals

The present study complied with the Helsinki-II declaration and was approved by the Danish Data Protection Agency, which provided a waiver for access to patient files without informed consent. Retrospective studies do not require ethical approval in accordance with national legislation.

Results

Patients

A total of 105 patients were identified from the hospital records, of whom 55 patients with 63 BS pairs (cases)

Table 2 Patient demographics and baseline characteristics

Total patient cases ($N = 63$)	
Age [mean (range)] (years)	70 (53–89)
Baseline PSA [median (range)] (ng/ml)	253 (4.3–9.708)
Disease stage [*n* (%)]	
Hormone sensitive	8 (12.7)
mHormone sensitive	19 (30.2)
CRPC	1 (1.6)
mCRPC	35 (55.6)
Treatment [*n* (%)]	
ADT	26 (41.3)
NGH before chemotherapy	8 (12.7)
NGH after chemotherapy	3 (4.8)
Chemotherapy	26 (41.3)

ADT, androgen-deprivation therapy; CRPC, castration-resistant prostate cancer; mCRPC, metastatic castration-resistant prostate cancer; mHormone sensitive, metastatic hormone sensitive; NGH, next-generation hormonal therapy; PSA, prostate-specific antigen.

were included in the final analysis (Table 2). Eight patients provided more than one dataset. Fifty patients were excluded for the following reasons: (a) the baseline and follow-up BS were not related to the same treatment regimen ($n=38$), (b) follow-up BS was performed less than 12 weeks from baseline ($n=7$), and (c) patient files or images were not retrievable ($n=5$). Baseline BS was performed from 68 days before the initiation of treatment to 12 days after treatment initiation. Follow-up BS was performed from 12 to 48 weeks within treatment with ADT (median: 26 weeks) and 12–28 weeks within treatment with NGH and chemotherapy (median: 18 weeks).

Biochemical and bone scan responses

All cases were classified by prostate-specific antigen (PSA) response according to PCWG-2 criteria [10], except for responders who were classified according to the PCWG-1 criteria, in which the response is classified as PSA decrease by 50% or more from baseline to follow-up [18]; this classification is not included in PCWG-2. The distribution of PSA response was as follows: (a) 28 responders, (b) eight with stable PSA from baseline to follow-up, (c) 22 so-called drifters, that is, initial PSA response, followed by a slow increase, and (d) five with PD, that is, more than or equal to 25% increase in PSA from baseline to follow-up.

A total of 54 (86%) out of the 63 cases presented with metastatic disease at baseline. Overall, progression from baseline was observed in 33% of the cases by standard clinical assessment, 35% of the cases by the MD Anderson criteria, and 29% of the cases by the PCWG-2 criteria (Table 3).

Observer agreement on response assessment

Observer agreement for standard clinical assessment was moderate (Table 4). The MD Anderson criteria showed moderate agreement and the PCWG-2 criteria showed considerable to almost perfect agreement (Table 4). Linear weighting of the MD Anderson criteria did not alter the κ values significantly ($\kappa=0.64$ vs. 0.60 when unweighted), with considerably overlapping CIs.

Cohen's κ is per se affected by the number of response categories, that is, the more the categories, the lower the κ values. Thus, we recalculated observer agreement for the standard clinical assessment and the MD Anderson on the basis of a reorganization of responses into PD and non-PD; stable disease and regression comprised the non-PD category. Observer agreement for the MD Anderson criteria increased from moderate ($\kappa=0.60$, Table 4) to substantial ($\kappa=0.76$, 95% CI: 0.58–0.93). A dichotomous response classification of standard clinical assessment did not alter the agreement ($\kappa=0.53$ vs. 0.52 with the original categories, with considerably overlapping CIs).

Table 3 Bone scan responses per assessment method [*n* (%)]

	Progressive disease	Stable disease	Response	Complete response
Standard clinical assessment	21 (33.3)[a]	34 (54.0)	8 (12.7)	–
MD Anderson	22 (34.9)	29 (46.0)	12 (19.0)	0 (0.0)
PCWG-2	18 (28.6)	45 (71.4)	–	–

PCWG-2, Prostate Cancer Working Group 2.
[a]Includes eight cases of 'mixed response'.

Table 4 Agreement in response assessment by response assessment method

Response assessment method	Unweighted Cohen's κ (95% CI)
Standard clinical assessment	0.52 (0.36–0.69)
MD Anderson	0.60 (0.44–0.77)
PCWG-2	0.84 (0.69–0.99)

CI, confidence interval; PCWG-2, Prostate Cancer Working Group 2.

Final consensus for response assessment

In 50 (79%) of the 63 cases, the final consensus on non-PD versus PD was the same across the three assessment methods. However, in more than 20% of the cases, the final consensus for response varied across the three different assessment methods, that is, one or two methods resulted in non-PD, whereas the other(s) resulted in PD. An illustrative example is shown in Fig. 1.

Observer agreement on M-status at baseline

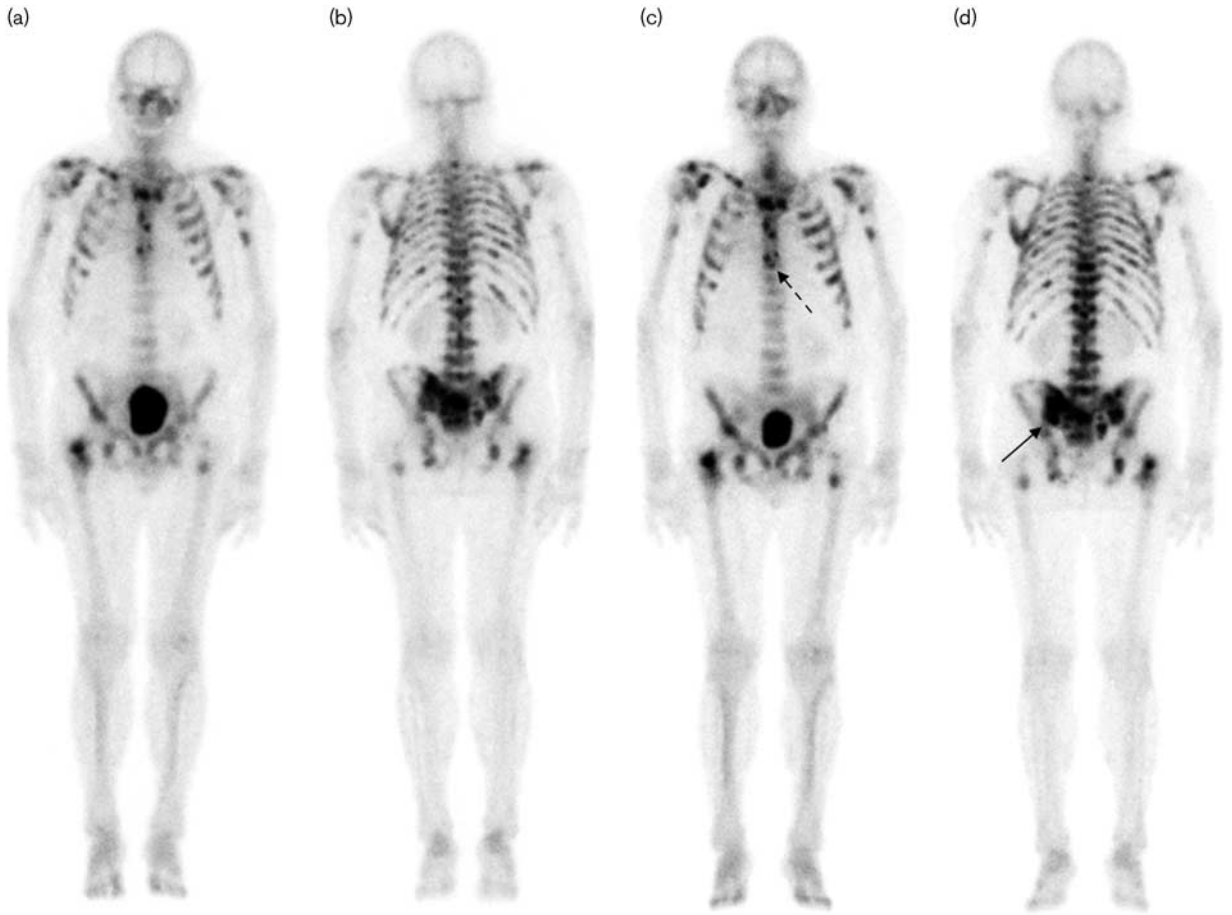
The variation in the response assessment could reflect a general variation in image reading. However, it was documented that the readers consistently classified the presence or absence of bone metastasis at baseline. These data showed almost perfect agreement ($\kappa=0.94$, 95% CI: 0.82–1.00). Similarly, considerable agreement was found for the assessment of super scan at baseline, although in this case, CIs were wider ($\kappa=0.78$, 95% CI: 0.49–1.00).

Lesion count on individual images

The Bland–Altman analysis showed that the mean difference in lesion count between readers was low, ranging from -0.8 to 0.5 at baseline (Table 5). This means that no reader systematically assessed the number of bone metastases higher than the others. However, a large variation in lesion count at the patient level was observed with limits of agreement from -15 to 15. The lowest variation was observed in the skull region, followed by the extremities, thorax, pelvis, and the spine, which showed the largest variation (Table 5). The skull and extremities showed the lowest number of bone metastases and the lowest limits of agreement, indicating that the lower the number of bone metastases, the lower the variation. This is also evident from the Bland–Altman plots for the pelvis, thorax, and spine (Fig. 2).

Subsequent categorization of patients according to the total lesion count classification, EOD, as defined by Soloway [14], showed almost perfect agreement by linear-weighted κ (0.87, 95% CI: 0.80–0.95).

Fig. 1



Patient with different outcomes when using the three assessment methods. This patient presented with numerous bone metastases at baseline; (a) anterior and (b) posterior. He received four cycles of chemotherapy (docetaxel) and had a follow-up bone scintigraphy at week 8; (c) anterior and (d) posterior. By standard clinical assessment and MD Anderson criteria, readers agreed upon 'stable disease'. The readers agreed on a decrease in intensity in several areas of the follow-up bone scan, including the thorax, the columna, and the pelvic area, but agreed that the general appearance of the scan was unchanged. According to the Prostate Cancer Working Group 2 criteria, progression is defined by the presence of two or more new lesions. One reader identified only one new lesion, dotted arrow (c), and thus deemed the response as nonprogression, whereas the other reader identified two new lesions, full arrow (d) and dotted arrow (c); they reached the consensus that two new lesions were present, and thus that there indeed was progression.

Table 5 Regional lesion count at baseline

Region	Median	Range	Mean difference	Lower limit of agreement	Upper limit of agreement
Skull	0	0–16	0.5	−4	5
Thorax	10	0 to > 20	−0.8	−14	13
Spine	8	0 to > 20	0.4	−15	15
Pelvis	6	0 to > 20	0.1	−13	13
Extremities	3	0 to > 20	0.2	−9	10

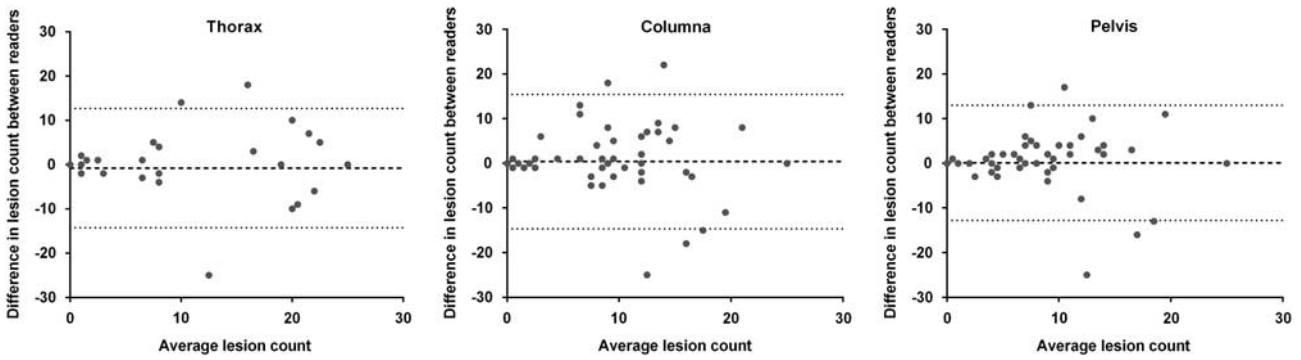
Mean difference, mean difference in lesion count between readers.

Discussion

Correct classification of BS and therefore agreement in interpretation of treatment effect is the key to patient management. Misclassification might deprive patients of beneficial treatments or result in patients receiving costly, noneffective treatments that might lead to severe side effects. Despite the fact that BS is used widely, as it is the recommended imaging modality for the diagnosis and monitoring of bone metastases, very few studies have investigated the observer agreement for evaluating treatment response on BS. From our results, we may conclude that the PCWG-2 criteria provided excellent agreement for the assessment of disease progression.

Radiographic progression is the key in the management of patients with advanced prostate cancer. At first glance, the methods applied performed similarly by classifying ~30% of the patients with PD. However, the level of agreement between readers for evaluating bone response during treatment depended notably on the method or the criteria applied. Several factors could explain the variance among readers in the response assessment. The disagreement in response assessment did not reflect a general variation in the reading of bone scans because classification of M1/M0 disease at baseline showed consistent results. The measures of observer agreement are often reported with crude

Fig. 2



Bland–Altman plot to evaluate variation in regional lesion count for the thorax, the columna, and the pelvis. The mean difference in the number of bone metastases between readers was low (dashed line), range: −0.79 to 0.38. The upper and lower limits of agreement (dotted lines) ranged from −14 to 13 for the thorax, −15 to 15 for the columna, and −13 to 13 for the pelvic region.

agreement without taking into consideration the number of classification categories or the prevalence of disease in the population. κ statistics take into consideration agreement by chance. Standard clinical assessment and the MD Anderson criteria both have four response categories, versus two categories with the PCWG-2 criteria, meaning that there is a higher probability of disagreement and consequently, agreement will be lower [15]. In the study by Zacho *et al.* [19], crude agreement for the classification of bone metastasis among three readers in 635 patients was 96% ($\kappa = 0.87$) for dichotomous M1/M0 classification, but only 66% ($\kappa = 0.57$) when reported on a four-point scale. However, recategorization of the standard clinical assessment and the MD Anderson criteria into PD versus non-PD did not change observer agreement with the standard clinical assessment. Such dichotomous reorganization of responses improved observer agreement with the MD Anderson criteria; however, agreement did not reach the level as seen with the PCWG-2 criteria. The interpretation of images by the standard clinical assessment and the MD Anderson criteria is more subjective compared with the PCWG-2 criteria. Interpretation of response by these methods was entirely at the discretion of the individual reviewer and their experience and based on subjective interpretation of images without exact criteria for change [9]. This may have caused some variation in the response assessment among readers. No intraobserver studies were carried out to clarify this hypothesis.

κ values are frequently reported without specification on the type of κ. Unweighted κ is mostly used. Weighted κ can be performed if there is a rank of categories among response groups [19]. For the unweighted κ, the relative importance of disagreement between categories is the same for adjacent categories as it is for distant categories, whereas for the weighted κ, the disagreements are ranked according to the relative distance between the categories [15]. Thus, with weighted κ, the observer agreement with the MD Anderson criteria would be lower if one reader reported complete regression rather than partial regression and the other reader reported stable disease. However, observer agreement was not improved by the use of weighted κ in our study.

The PCWG-2 criteria might be preferred on the basis of the high level of agreement between readers, even though valuable information may be lost when there is no distinction between disappearance and/or regression of lesions and stable disease. The choice of two or more new lesions might seem arbitrary; however, it has been shown that radiographic progression, according to PCWG-2, is associated significantly with overall survival [20]. In addition, the most important clinical distinction is the presence of PD or non-PD with respect to anticancer medication, which is continued if tolerated in stable and responding patients.

Similar to our results, Kaboteh *et al.* [11] showed good agreement between three readers for bone response assessment by PCWG-2 criteria in prostate cancer patients with and without bone metastases. However, no κ values were reported. In the sub-population of 173 patients with bone metastasis from the study by Kaboteh and colleagues, Anand *et al.* [12] showed κ values from 0.55 to 0.81 for observer agreement on BS response assessment by PCWG-2 criteria between three readers. They did not observe improved observer agreement with PCWG-2 versus an unspecified clinical assessment (increased burden of metastases). These data are in contrast to our data, which showed superior observer agreement with PCWG-2 for response assessment.

Observer variation can be influenced by many factors, including institutional standardization or habituation of disease classification and experience. In studies by Kaboteh and colleagues and Anand and colleagues, readers were recruited from the same institution. With respect to classification of metastases on bone scans, modest κ values have been reported using a four-point scale among physicians from 18 centers in Sweden

(κ=0.48) [21]. Our current data on M1/M0 classification paralleled the results from other studies [19,22]. Our study confirmed excellent overall agreement on treatment response assessment with the PCWG-2 criteria with five board-certified nuclear medicine physicians from different institutions and with varying experience with bone imaging, indicating the robustness of this response assessment method. Finally, the two BS were performed within the same treatment regimen, thus reflecting standard clinical practice where follow-up BS is used to assess disease status during treatment. Kaboteh and colleagues and Anand and colleagues included patients with two BS performed within a time period of 7 years, but not necessarily from within the same treatment.

Lesion counting is one way of reporting treatment outcome if images are to be assessed without bias from proceeding images or reports. Our findings indicated that the counting of individual lesions resulted in a large variation from reader to reader. The variation was the lowest when the number of bone metastases was low. These data imply high observer variation with increasing number of lesions. Thus, variation is high in skeletal regions most commonly affected by bone metastases in prostate cancer, namely, the spine and pelvis. The number of lesions in these regions is often high in advanced prostate cancer. Agreement improved notably when lesion numbers were reorganized according to the EOD classification by Soloway *et al.* [14], which has been associated with overall survival. Thus, this method might be preferred over simple counting.

Conclusion

The PCWG-2 criteria showed excellent agreement for BS response assessment, whereas less agreement was observed among experienced readers with other methods. Reproducible classification of bone status is important in clinical decision-making. We recommend the use of stringent and objective criteria, such as the PCWG-2 criteria, both in clinical trials and in clinical practice. The separate counting of lesions without access to previous scans cannot be recommended. If lesion counting is required, we recommend the use of EOD grades and a side-by-side comparison of BS.

Acknowledgements

Authors' contribution: R.F.F., H.D.Z., and L.J.P. conceived the study; R.F.F., H.D.Z., and L.J.P. participated in its design and coordination; R.F.F. and S.A. performed data collection; H.D.Z., J.F., J.A.E., M.H.C., and R.A. read and evaluated bone scintigraphies; J.B.M. carried out all statistical analyses; R.F.F., H.D.Z., and L.J.P. drafted the manuscript. All authors critically revised the manuscript and approved the final version to be published.

This study was supported by the Obel Family Foundation.

Conflicts of interest

There are no conflicts of interest.

References

1 Siegel RL, Miller KD, Jemal A. Cancer statistics 2016. *CA Cancer J Clin* 2016; **66**:7–30.
2 Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, *et al.* Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* 2013; **49**:1374–1403.
3 Savage P, Mahmoud S. Development and economic trends in cancer therapeutic drugs: a 5-year update 2010–2014. *Br J Cancer* 2015; **112**:1037–1041.
4 Pilon D, Queener M, Lefebvre P, Ellis LA. Cost per median overall survival month associated with abiraterone acetate and enzalutamide for treatment of patients with metastatic castration-resistant prostate cancer. *J Med Econ* 2016; **19**:777–784.
5 Sanyal C, Aprikian AG, Cury FL, Chevalier S, Dragomir A. Management of localized and advanced prostate cancer in Canada: a lifetime cost and quality-adjusted life-year analysis. *Cancer* 2016; **122**:1085–1096.
6 Heidenreich A, Bastian PJ, Bellmunt J, Bolla M, Joniau S, van der Kwast T, *et al.* EAU guidelines on prostate cancer. Part II: treatment of advanced, relapsing, and castration-resistant prostate cancer. *Eur Urol* 2014; **65**:467–479.
7 National Comprehensive Cancer Network. NCCN guidelines, version 1.2016. Prostate cancer; 2016. Available at: https://www.nccn.org/professionals/physician_gls/pdf/prostate.pdf. [Accessed 13 January 2016].
8 Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, *et al.* New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst* 2000; **92**:205–216.
9 Hamaoka T, Madewell JE, Podoloff DA, Hortobagyi GN, Ueno NT. Bone imaging in metastatic breast cancer. *J Clin Oncol* 2004; **22**:2942–2953.
10 Scher HI, Halabi S, Tannock I, Morris M, Sternberg CN, Carducci MA, *et al.* Design and end points of clinical trials for patients with progressive prostate cancer and castrate levels of testosterone: recommendations of the Prostate Cancer Clinical Trials Working Group. *J Clin Oncol* 2008; **26**:1148–1159.
11 Kaboteh R, Gjertsson P, Leek H, Lomsky M, Ohlsson M, Sjostrand K, *et al.* Progression of bone metastases in patients with prostate cancer – automated detection of new lesions and calculation of bone scan index. *EJNMMI Res* 2013; **3**:64.
12 Anand A, Morris MJ, Kaboteh R, Bath L, Sadik M, Gjertsson P, *et al.* Analytic validation of the automated bone scan index as an imaging biomarker to standardize quantitative changes in bone scans of patients with metastatic prostate cancer. *J Nucl Med* 2016; **57**:41–45.
13 Bombardieri E, Aktolun C, Baum RP, Bishof-Delaloye A, Buscombe J, Chatal JF, *et al.* Bone scintigraphy: procedure guidelines for tumour imaging. *Eur J Nucl Med Mol Imaging* 2003; **30**:99–106.
14 Soloway MS, Hardeman SW, Hickey D, Raymond J, Todd B, Soloway S, *et al.* Stratification of patients with metastatic prostate cancer based on extent of disease on initial bone scan. *Cancer* 1988; **61**:195–202.
15 Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003; **228**:303–308.
16 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–174.
17 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**:135–160.
18 Bubley GJ, Carducci M, Dahut W, Dawson N, Daliani D, Eisenberger M, *et al.* Eligibility and response guidelines for phase II clinical trials in androgen-independent prostate cancer: recommendations from the Prostate-Specific Antigen Working Group. *J Clin Oncol* 1999; **17**:3461–3467.
19 Zacho HD, Manresa JAB, Mortensen JC, Bertelsen H, Petersen LJ. Observer agreement and accuracy in the evaluation of bone scans in newly diagnosed prostate cancer. *Nucl Med Commun* 2015; **36**:445–451.
20 Sonpavde G, Pond GR, Armstrong AJ, Galsky MD, Leopold L, Wood BA, *et al.* Radiographic progression by Prostate Cancer Working Group (PCWG)-2 criteria as an intermediate endpoint for drug development in metastatic castration-resistant prostate cancer. *BJU Int* 2014; **114**: E25–E31.
21 Sadik M, Suurkula M, Hoglund P, Jarund A, Edenbrandt L. Quality of planar whole-body bone scan interpretations – a nationwide survey. *Eur J Nucl Med Mol Imaging* 2008; **35**:1464–1472.
22 Ore L, Hardoff R, Gips S, Tamir A, Epstein L. Observer variation in the interpretation of bone scintigraphy. *J Clin Epidemiol* 1996; **49**:67–71.