

3. Adams HJ, Kwee TC. Will treatment intensification in early-stage Hodgkin lymphoma patients with a positive interim FDG-PET improve outcome? *Pediatr Hematol Oncol*. 2016;33:1–4.
4. Andre MP, Girinsky T, Federico M, et al. Early Positron Emission Tomography Response-Adapted Treatment in Stage I and II Hodgkin Lymphoma: Final Results of the Randomized EORTC/LYSA/FIL H10 Trial. *J Clin Oncol*. 2017; JCO2016686394.
5. Raemaekers JM, André MP, Federico M, et al. Omitting radiotherapy in early positron emission tomography-negative stage I/II Hodgkin lymphoma is associated with an increased risk of early relapse: clinical results of the preplanned interim analysis of the randomized EORTC/LYSA/FIL H10 trial. *J Clin Oncol*. 2014;32:1188–1194.
6. Radford J, Illidge T, Counsell N, et al. Results of a trial of PET-directed therapy for early-stage Hodgkin's lymphoma. *N Engl J Med*. 2015;372:1598–1607.
7. Adams HJ, Nievelstein RA, Kwee TC. Prognostic value of interim FDG-PET in Hodgkin lymphoma: systematic review and meta-analysis. *Br J Haematol*. 2015;170:356–366.
8. Engert A, Plutschow A, Eich HT, et al. Reduced treatment intensity in patients with early-stage Hodgkin's lymphoma. *N Engl J Med*. 2010;363:640–652.
9. Adams HJ, Kwee TC. A negative ¹⁸F-FDG-PET scan can never exclude residual disease. *Nucl Med Commun*. 2016;37:102–103.
10. Adams HJ, Kwee TC. Negative PET: no guarantee of good prognosis in Hodgkin's lymphoma. *Ann Hematol*. 2015;94:1609–1610.
11. Klimm B, Goergen H, Fuchs M, et al. Impact of risk factors on outcomes in early-stage Hodgkin's lymphoma: an analysis of international staging definitions. *Ann Oncol*. 2013;24:3070–3076.
12. Gallamini A, Hutchings M, Ritacco L, et al. Early interim 2-[¹⁸F]fluoro-2-deoxy-D-glucose positron emission tomography is prognostically superior to international prognostic score in advanced-stage Hodgkin's lymphoma: a report from a joint Italian-Danish study. *J Clin Oncol*. 2007;25:3746–3752.
13. Gallamini A, Barrington SF, Biggi A, et al. The predictive role of interim positron emission tomography for Hodgkin lymphoma treatment outcome is confirmed using the interpretation criteria of the Deauville five-point scale. *Haematologica*. 2014;99:1107–1113.
14. Agostinelli C, Gallamini A, Stracqualursi L, et al. The combined role of biomarkers and interim PET scan in prediction of treatment outcome in classical Hodgkin's lymphoma: a retrospective, European, multicentre cohort study. *Lancet Haematol*. 2016;3:e467–e479.
15. Adams HJ, Kwee TC. Prevention of large-scale implementation of unnecessary and expensive predictive tests in Hodgkin's lymphoma. *Lancet Haematol*. 2017;4:e63–e64.
16. Adams HJ, Kwee TC. Proportion of false-positive lesions at interim and end-of-treatment FDG-PET in lymphoma as determined by histology: systematic review and meta-analysis. *Eur J Radiol*. 2016;85:1963–1970.
17. Avivi I, Zilberlicht A, Dann EJ, et al. Strikingly high false positivity of surveillance FDG-PET/CT scanning among patients with diffuse large cell lymphoma in the rituximab era. *Am J Hematol*. 2013;88:400–405.
18. Nakayama H, Aisa Y, Ito C, et al. Importance of histologic verification of positive positron emission tomography/computed tomography findings in the follow-up of patients with malignant lymphoma. *Clin Lymphoma Myeloma Leuk*. 2015;15:753–760.
19. Zinzani PL, Tani M, Trisolini R, et al. Histological verification of positive positron emission tomography findings in the follow-up of patients with mediastinal lymphoma. *Haematologica*. 2007;92:771–777.
20. El-Galaly TC, Mylam KJ, Brown P, et al. Positron emission tomography/computed tomography surveillance in patients with Hodgkin lymphoma in first remission has a low positive predictive value and high costs. *Haematologica*. 2012;97:931–936.
21. Adams HJ, Kwee TC. Controversies on the prognostic value of interim FDG-PET in advanced-stage Hodgkin lymphoma. *Eur J Haematol*. 2016;97:491–498.
22. Bakhshi S, Bhethanabhotla S, Kumar R, et al. Posttreatment PET/CT rather than interim PET/CT using Deauville criteria predicts outcome in pediatric Hodgkin lymphoma: a prospective study comparing PET/CT with conventional imaging. *J Nucl Med*. 2017;58:577–583.
23. Mesguich C, Cazeau AL, Bouabdallah K, et al. Hodgkin lymphoma: a negative interim-PET cannot circumvent the need for end-of-treatment-PET evaluation. *Br J Haematol*. 2016;175:652–660.
24. Johnson P, Federico M, Kirkwood A, et al. Adapted treatment guided by interim PET-CT scan in advanced Hodgkin's lymphoma. *N Engl J Med*. 2016;374:2419–2429.
25. Press OW, Li H, Schoder H, et al. US intergroup trial of response-adapted therapy for stage III to IV Hodgkin lymphoma using early interim fluorodeoxyglucose-positron emission tomography imaging: Southwest Oncology Group S0816. *J Clin Oncol*. 2016;34:2020–2027.
26. Zinzani PL, Broccoli A, Gioia DM, et al. Interim positron emission tomography response-adapted therapy in advanced-stage Hodgkin lymphoma: final results of the phase II part of the HD0801 study. *J Clin Oncol*. 2016;34:1376–1385.
27. Adams HJ, Kwee TC. Interim PET-CT scan in advanced Hodgkin's lymphoma [comment]. *N Engl J Med*. 2016;375:999.
28. Adams HJ, Kwee TC. Predictive value of interim [¹⁸F]fluorodeoxyglucose-positron emission tomography in advanced-stage Hodgkin lymphoma is not well established. *J Clin Oncol*. 2017;35:370–371.
29. Adams HJ, Kwee TC. Does interim ¹⁸F-FDG-PET response-adapted therapy really benefit advanced-stage Hodgkin lymphoma patients? *Nucl Med Commun*. 2016;37:1333–1334.
30. Casasnovas O, Brice P, Bouabdallah R, et al. Randomized phase III study comparing an early PET driven treatment de-escalation to a not PET-monitored strategy in patients with advanced stages Hodgkin lymphoma: interim analysis of the AHL2011 Lysa study [abstract]. *Blood*. 2015;126:577.

Hugo J.A. Adams*

Thomas C. Kwee

*Deventer Ziekenhuis

Nico Bolkesteinlaan 75

7416 SE Deventer, The Netherlands

E-mail: h.j.a.adams@gmail.com

Published online Mar. 9, 2017.

DOI: 10.2967/jnumed.117.192294

A Common Mistake in Assessing the Diagnostic Value of a Test: Failure to Account for Statistical and Methodologic Issues

TO THE EDITOR: I was interested to read the paper by Anand et al. in the December 2016 edition of *The Journal of Nuclear Medicine* (1). The purpose of the authors was to assess the impact of variability in scanning speed and in vendor-specific γ -camera settings on the reproducibility and accuracy of the automated bone scan index (BSI) (1). They measured reproducibility as the absolute difference between repeated BSI values, and they measured accuracy as the absolute difference between observed BSI values and phantom BSI values. Descriptive statistics were used to compare the generated data. Reproducibility (reliability) and accuracy (validity), as two completely different methodologic issues, should be assessed using appropriate tests. It is crucial to be aware that, regarding reliability, one should use the intraclass correlation coefficient for quantitative variables and the weighted κ -test for qualitative variables. However, regarding validity, one should use the interclass correlation coefficient (Pearson r) for quantitative variables whereas the most appropriate tests for qualitative variables may include sensitivity, specificity, positive and negative predictive value, positive and negative likelihood ratio, diagnostic accuracy, and odds ratio. Moreover, in analyzing reliability, one should apply an individual-based approach using single-measure intraclass correlation coefficient agreement because applying a global-average approach (absolute difference) can be misleading. A test may indicate high validity, yet there may be no reliability at all (2–8).

Anand et al. enrolled 25 patients in each of 3 groups and observed a significantly lower reproducibility for group 2 (mean \pm SD, 0.35 ± 0.59) than for group 1 (0.10 ± 0.13 ; $P < 0.0001$) or group 3 (0.09 ± 0.10 ; $P < 0.0001$). However, no significant difference in reproducibility was observed between group 3 and group 1 ($P = 0.388$) (1). Statistical significance and clinical

importance are two completely different issues, and in clinical research—especially reliability analysis—we should not put the emphasis on significance level (P value) (2–8).

They concluded that the accuracy and reproducibility of automated BSI were dependent on scanning speed but not on vendor-specific γ -camera settings. Such a conclusion should be supported by the above-mentioned statistical and methodologic issues. Otherwise, in clinical practice, misdiagnosis and patient mismanagement may occur.

REFERENCES

1. Anand A, Morris MJ, Kaboteh R, et al. A preanalytic validation study of automated bone scan index: effect on accuracy and reproducibility due to the procedural variabilities in bone scan image acquisition. *J Nucl Med*. 2016;57:1865–1871.
2. Szklo M, Nieto FJ. *Epidemiology Beyond the Basics*. 2nd ed. Manhattan, NY: Jones and Bartlett; 2007.
3. Sabour S. Myocardial blood flow quantification by Rb-82 cardiac PET/CT: methodological issues on reproducibility study. *J Nucl Cardiol*. September 6, 2016 [Epub ahead of print].
4. Sabour S. Reproducibility of semi-automatic coronary plaque quantification in coronary CT angiography with sub-mSv radiation dose: common mistakes. *J Cardiovasc Comput Tomogr*. 2016;10:e21–e22.
5. Sabour S. Reliability of a new modified tear breakup time method: methodological and statistical issues. *Graefes Arch Clin Exp Ophthalmol*. 2016;254:595–596.
6. Sabour S, Farzaneh F, Peymani P. Evaluation of the sensitivity and reliability of primary rainbow trout hepatocyte vitellogenin expression as a screening assay for estrogen mimics: methodological issues. *Aquat Toxicol*. 2015;164:175–176.
7. Sabour S. Re: does the experience level of the radiologist, assessment in consensus, or the addition of the abduction and external rotation view improve the diagnostic reproducibility and accuracy of MRA of the shoulder? [comment]. *Clin Radiol*. 2015;70:333–334.
8. Sabour S. The reliability of routine clinical post-processing software in assessing potential diffusion-weighted MRI “biomarkers” in brain metastases, common mistake [comment]. *Magn Reson Imaging*. 2014;32:1162.

Siamak Sabour

Shahid Beheshti University of Medical Sciences
Tehran 1413663633, Islamic Republic of Iran
E-mail: s.sabour@sbmu.ac.ir

Published online Jan. 12, 2017.
DOI: 10.2967/jnumed.115.156745

REPLY: Thank you for your interest and comments regarding our work (1). We agree that reproducibility/reliability and accuracy/validity are two different methodologic issues and should be evaluated using appropriate tests for each. Further, we also concur that for quantitative variables, the intraclass correlation coefficient should be used to test reliability and the interclass correlation coefficient (Pearson r) should be used to test validity. Since the automated bone scan index (BSI) is a quantitative parameter, the comments on qualitative assessments are inapplicable.

Our analytic validation study demonstrated the validity of the automated BSI with a Pearson r of 0.99 ($P < 0.0001$) and with associated parameters of linear regression (slope, 0.80 [95% confidence interval, 0.78–0.83]; intercept, 0.38 [95% confidence interval, 0.25–0.51]) (2). In this study, Cohen κ -agreement and other well-known standard tests for imaging biomarkers (3) were also used to evaluate the reliability of the automated BSI platform. These tests demonstrated a coefficient of variation (<20%) with a consistent linearity—satisfying

the homoscedasticity and Shapiro–Wilken test—from low-burden to high-burden disease. The Bland–Altman plot of absolute difference in BSI readings of repeated scans revealed the analytic noise with an SD of 0.15. The study concluded that in the clinical setting, the assessment of change in BSI above the noise threshold (2 SDs, or 0.30) had high interobserver agreement (Cohen $\kappa = 0.96$).

Despite the analytic validation of BSI, variation in the imaging procedure can significantly affect the analytic performance of the imaging biomarker. Therefore, in a subsequent publication, we analyzed the effect of variability in scanning speed and γ -camera settings on the BSI assessment (1). The objective of our preanalytic study was not to evaluate the performance characteristics of BSI but to use appropriate statistical tests to evaluate the effect of procedural variability on the analytically validated BSI values. The Wilcoxon/Mann–Whitney test (4) demonstrated that the noise in the BSI value was significantly higher because of variability in scanning speed in comparison with the known analytic gold standard (accuracy performance) and with repeated measurement of test–retest patient bone scans (reproducibility performance). The P value in our study, from the Wilcoxon/Mann–Whitney test, did not imply clinical significance but referred to the statistical significance of the degree of systematic difference between the two BSI readings of the repeated bone scans that were performed in a specific study design to assess the effect of procedural variability on the known performance characteristics of BSI.

In conclusion, the statistical methodology must be reviewed in the context of the study design and its objective. We have performed a comprehensive assessment to demonstrate the reliability and validity of BSI, and with the preanalytic study, we demonstrated that the BSI performance characteristics were dependent on scanning speed. The study added empiric evidence toward the standardization of bone scan acquisitions for robust quantitative BSI assessment in multiinstitutional studies. Together, the analytic and preanalytic studies served as the foundation for prospective clinical investigations aimed at validating automated BSI as a quantitative imaging biomarker.

REFERENCES

1. Anand A, Morris MJ, Kaboteh R, et al. A preanalytic validation study of automated bone scan index: effect on accuracy and reproducibility due to the procedural variabilities in bone scan image acquisition. *J Nucl Med*. 2016;57:1865–1871.
2. Anand A, Morris MJ, Kaboteh R, 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.
3. O'Connor JP, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*. 2017;14:169–186.
4. Szklo M, Nieto FJ. *Epidemiology Beyond the Basics*. 2nd ed. Manhattan, NY: Jones and Bartlett; 2007.

Aseem U. Anand*
David Minarik

*Lund University
Waldenströms Gata 5, Second Floor
Malmö 20502, Sweden
E-mail: aseem.anand@med.lu.se

Published online Apr. 6, 2017.
DOI: 10.2967/jnumed.117.190645