Endpoint comparison for bone mineral density measurements in North Central Cancer Treatment Group cancer clinical trials N02C1 and N03CC (Alliance)

Alliance Clinical Trials Oncology

Citation of published version:
Dueck, A. C., Singh, J., Atherton, P., Liu, H., Novotny, P., Hines, S., Loprinzi, C. L., Perez, E. A., Tan, A., Burger, K., Zhao, X., Diekmann, B., & Sloan, J. A. (2015). Endpoint comparison for bone mineral density measurements in North Central Cancer Treatment Group cancer clinical trials N02C1 and N03CC (Alliance). In Osteoporosis International (Vol. 26, Issue 7, pp. 1971–1977). Springer Science and Business Media LLC. https://doi.org/10.1007/s00198-015-3091-4

HHS Public Access

Author manuscript

Osteoporos Int. Author manuscript; available in PMC 2016 July 01.

Published in final edited form as:

Osteoporos Int. 2015 July ; 26(7): 1971–1977. doi:10.1007/s00198-015-3091-4.

Endpoint comparison for bone mineral density measurements in North Central Cancer Treatment Group cancer clinical trials N02C1 and N03CC (Alliance)

A. C. Dueck,
Alliance Statistics and Data Center, Division of Health Sciences Research, Mayo Clinic, 13400 E. Shea Blvd., Scottsdale, AZ 85259, USA

J. Singh,
Department of Medicine, University of Alabama, Tuscaloosa, AL, USA

P. Atherton,
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

H. Liu,
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

P. Novotny,
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

S. Hines,
Department of Radiation Oncology, Mayo Clinic, Rochester, MN, USA

C. L. Loprinzi,
Department of Medical Oncology, Mayo Clinic, Rochester, MN, USA

E. A. Perez,
Department of Oncology, Mayo Clinic, Jacksonville, FL, USA

A. Tan,
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

K. Burger,
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

X. Zhao,
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Correspondence to: A. C. Dueck, dueck@mayo.edu.

Conflicts of interest: None.

B. Diekmann, and
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

J. A. Sloan
Alliance Statistics and Data Center, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

for the Alliance for Clinical Trials in Oncology
A. C. Dueck: dueck@mayo.edu

Abstract

Summary—Bone mineral density (BMD) measurement can vary depending upon anatomical site, machine, and normative values used. This analysis compared different BMD endpoints in two clinical trials. Trial results differed across endpoints. Future clinical trials should consider inclusion of multiple endpoints in sensitivity analysis to ensure sound overall study conclusions.

Introduction—Methodological issues hamper efficacy assessment of osteoporosis prevention agents in cancer survivors. Osteoporosis diagnosis can vary depending upon which bone mineral density (BMD) anatomical site and machine is used and which set of normative values are applied. This analysis compared different endpoints for osteoporosis treatment efficacy assessment in two clinical studies.

Methods—Data from North Central Cancer Treatment Group phase III clinical trials N02C1 and N03CC (Alliance) were employed involving 774 patients each comparing two treatments for osteoporosis prevention. Endpoints for three anatomical sites included raw BMD score (RawBMD); raw machine-based, sample-standardized, and reference population-standardized T scores (RawT, TSamp, TRef); and standard normal percentile corresponding to the reference population-standardized T score (TPerc). For each, treatment arm comparison was carried out using three statistical tests using change and percentage change from baseline (CB, %CB) at 1 year.

Results—Baseline correlations among endpoints ranged from 0.79 to 1.00. RawBMD and TPerc produced more statistically significant results (14 and 19 each out of 36 tests) compared to RawT (11/36), TSamp (8/36), and TRef (7/36). Spine produced the most statistically significant results (26/60) relative to femoral neck (20/60) and total hip (13/60). Lastly, CB resulted in 44 statistically significant results out of 90 tests, whereas %CB resulted in only 15 significant results.

Conclusions—Treatment comparisons and interpretations were different across endpoints and anatomical sites. Transforming via sample statistics provided similar results as transforming via reference or machine-based norms. However, RawBMD and TPerc may be more sensitive to change as clinical trial endpoints.

Keywords

Bone mineral density; Breast cancer; Cancer survivorship; Endpoints; Osteoporosis

Introduction

Osteoporosis is a significant public health problem due to its impact on morbidity and mortality in the elderly [1–3]. In recent years, there have been numerous attempts to discover and validate treatments for the prevention and amelioration of osteoporosis [4]. Most recently, a concern was raised about the potential increase in fractures resulting from antiosteoporosis treatments in breast cancer patients that turned out to be untrue based on a large meta-analysis [5]. The North Central Cancer Treatment Group (NCCTG) completed clinical trials N02C1 and N03CC (Alliance) intended to assess interventions to prevent and/or treat osteoporosis in cancer patients.

Alongside these clinical investigations, there have been challenges and advances in defining and optimizing ways to assess the presence and severity of osteoporosis. Little foundational work has been performed to assess the relative merit of alternative measures for osteoporosis. This spurred the present investigation regarding how to best approach the analysis of these studies in a consistent manner. This manuscript reports results of a comparison of alternative computational approaches for bone mineral density (BMD) endpoints using data from N02C1 and N03CC. We propose a new computational method that may have some advantages over existing methods for assessing BMD in clinical trials.

The present gold standard for diagnosing osteoporosis is based on the World Health Organization (WHO) and International Society for Clinical Densitometry (ISCD) criteria [6]. In an attempt to attain some consistency across populations and sites, the identification of osteoporosis is based on whether an individual's BMD is sufficiently low relative to what it would be for a young healthy individual [7]. There is consensus that a T score below −2.5 (i.e., a person whose BMD is 2.5 standard deviations below mean BMD for a healthy individual) indicates osteoporosis [8]. AT score below −1 but not below −2.5 is indicative of osteopenia [9].

There are numerous sources of measurement error and considerable inconsistency in the way that BMD data are presented and analyzed in the literature [10]. The diagnosis of osteoporosis for an individual can vary depending upon which BMD site is used [11–16], which machine is used for the assessment, and which set of normative values are applied [17].

The T score is an often-used value intended to standardize BMD scores across population variability. One problem in the application of this value is defining which population should be used as the reference for calculation of the T scores. Table 1 presents a variety of reference data for BMD achieved through a local reference population or the machine manufacturers' values. The variety, both across machine and anatomical site, is considerable. The reference values hence can be different depending upon the anchor distribution used. This is consistent with other aspects of measurement such as assessing clinical significance of patient-reported outcomes [18]. Using WHO data versus other normative data can produce widely discrepant estimates for the prevalence of osteoporosis (a T score below −2.5) [19]. Studies assessing wide variety of endpoints for BMD measures in premenopausal and postmenopausal women show marked variability depending upon the

endpoint itself and even within the same endpoint for different reference populations [20]. Hence, there is a need for examination of the relative impact of alternative BMD endpoints on the results of clinical studies. The present study was intended to address this knowledge gap by carrying out a statistical comparison of alternative post hoc computational approaches to BMD measurement.

Methods

Patient population

We used data from two NCCTG (Alliance) clinical trials, namely N02C1 [21] and N03CC [22]. N02C1 was a phase III randomized, placebo-controlled, double-blind trial of risedronate for prevention of bone loss in premenopausal women undergoing chemotherapy for primary breast carcinoma. N03CC was a randomized, controlled, open-label trial of empiric prophylactic versus delayed use of zoledronic acid for prevention of bone loss in post-menopausal women with breast cancer initiating therapy with letrozole after tamoxifen. Study design, monitoring, treatment, and eligibility criteria for each clinical trial are presented elsewhere [21, 22]. Each participant across studies signed an IRB-approved, protocol-specific informed consent in accordance with federal and institutional guidelines.

Bone mineral density measurements

For these studies, the total lumbar spine was used as the anatomical site for the primary BMD endpoint; however, both protocols specified multiple other anatomical sites within secondary endpoints. The primary endpoint in N02C1 was change from baseline at 1 year post-randomization in spine BMD, whereas N03CC employed percentage change from baseline. Both protocols employed a comparison between arms using a two-sample *t* test.

In these studies, the amount of data acquired per subject varied, ranging from a single value to a series of BMD measures for various anatomical sites. There were numerous possible endpoints that could be used per anatomical site as the basis for analysis:

1. The observed raw BMD values (RawBMD)
2. The observed T scores (using a machine-based reference sample; RawT)
3. The transformed T scores using the sample mean and sample standard deviation (TSamp)
4. The transformed T scores using the mean and standard deviation from a reference population such as the National Health and Nutrition Examination Survey (NHANES) data [23], the WHO [24], or from the UK local reference population [10, 17] (TRef)
5. The standard normal percentile corresponding to the reference population-standardized [10] T score (TPerc)

RawBMD and RawT are bone mineral density values as reported on the case report forms, which are presumably the values reported via the bone mineral density measurement clinical findings report. To compute TSamp, we must first compute the sample mean and sample standard deviation of the baseline RawBMD values. Then, TSamp is computed by

subtracting the baseline sample mean from each RawBMD value and dividing by the baseline sample standard deviation. For example, for N02C1, the sample mean and sample standard deviation of the baseline spine RawBMD values was 1.21 and 0.16 (Table 2), respectively. So, the spine TSamp value would be computed for a patient with a spine RawBMD score of 1.13 at a given time point as (1.13–1.21)/0.16 or −0.50. TRef is similarly computed; however, a mean and standard deviation from a reference population is used instead of the baseline RawBMD sample mean and baseline RawBMD sample standard deviation. For example, using a reference population spine mean and standard deviation of 1.047 and 0.11 (Table 1), respectively, the spine TRef value would be computed for a patient with a spine RawBMD score of 1.13 at a given time point as (1.13–1.047)/0.11 or 0.75. Finally, TPerc is computed as the probability that an observation from the standard normal distribution is less than or equal to TRef. For the example in which a patient's TRef value was 0.75 at a given time point, TPerc is computed as $P(Z \leq 0.75)=0.77$.

All of the above endpoints, except for the percentile, which is a new idea, have been used in previous studies [8–10, 17, 20, 25–27]. However, there has been no discussion in the literature regarding the relative clinical significance of these various endpoints other than to indicate that one might make mistaken clinical conclusions depending upon which endpoint is used. It is unclear what each of these endpoints really means in terms of impact on data interpretation. All five endpoints have substantial inherent variability and measurement error. Note that TSamp and TRef are monotonic or order-preserving transformations in addition to being linear transformations of RawBMD, resulting in identical results when comparing absolute (but not percentage) change from baseline between groups using a Wilcoxon rank-sum test, a two-sample *t* test, or analysis of covariance (ANCOVA) for these three endpoints. These analyses are described in the following section. Data collection and statistical analyses were conducted by the Alliance Statistics and Data Center. All analyses were based on the frozen study databases reported in the primary manuscripts [21, 22].

Statistical analysis

Studies to date have been powered to detect various effect sizes, most often based on a percentage change using a Wilcoxon rank-sum test [23], two-sample *t* test [8], or ANCOVA procedure [26] comparing groups. The ANCOVA would typically include the corresponding baseline BMD endpoint value as a covariate.

For each of the five endpoints, change from baseline (CB) and percentage change from baseline (%CB) were calculated using all available data for eligible patients for three anatomical sites: spine (LTotal), femoral neck (femur neck), and total hip (femur total). Student's two-sample *t* test [28], Wilcoxon rank-sum test [29], and ANCOVA (corresponding baseline BMD endpoint value used as a covariate) [30] were used to compare treatment arms for each endpoint. Each clinical trial was analyzed separately. *P* values<0.05 were considered statistically significant throughout (i.e., as if the given analysis represented the primary comparison for the study).

To compare and contrast the various measures of BMD, we also calculated summary statistics and Pearson correlations [28] among the five endpoints at baseline.

Results

Details of the study populations are presented elsewhere [21, 22] but are summarized here. In total, 204 and 505 patients provided data for this analysis from N02C1 and N03CC, respectively. Patients ranged in age from 36 to 88 with a median age of 59 years. Ninety-seven percent of the patients were Caucasians.

Correlations among the five alternative osteoporosis endpoints at baseline ranged from 0.79 to 1.00 for each anatomical site within each study. RawBMD, TSamp, and TRef are perfectly correlated (i.e., are linear transformations of each other). Hence, all five endpoints are strongly correlated and some are pseudo-redundant psychometrically in terms of correlation. Table 2 provides descriptive statistics at baseline for each endpoint for spine bone mineral density.

The primary goal of this analysis was to see if the choice of endpoint would change the study outcome. Figures 1 and 2 present side-by-side boxplots of the five different endpoints based on change from baseline using spine measurements by treatment arm for each of the two studies. The statistical significance of this comparison differed depending upon the endpoint and test used within N02C1, but results were highly statistically significant for all endpoints and tests in N03CC.

We then compared the BMD analysis results when using different anatomical sites for generation of the BMD data (Table 3). We performed three statistical tests for each cell in the table comparing treatment arms (two-sample *t* test, Wilcoxon, ANCOVA). Among the five endpoints, RawT, TSamp, and TRef had 11, 8, and 7 statistically significant results out of 36 tests, respectively. RawBMD and TPerc produced the most statistically significant results, 14 and 19, respectively. Across the three anatomical sites, spine, femoral neck, and total hip had 26, 20, and 13 statistically significant results out of 60 tests, respectively. Lastly, in comparing CB to %CB, use of CB resulted in 44 statistically significant results out of 90 tests, whereas %CB results in only 15 significant results.

For N02C1 (reported as a negative study), we observed six statistically significant results overall out of 90 tests, with statistically significant results being limited to TPerc (5/18 tests) and RawT (1/18 tests). All other statistical results were non-significant. These statistically significant results were also limited to spine (5/30 tests) and femoral neck (1/30 tests) sites, with no differences observed for the total hip site regardless of endpoint or method. Percentage change from baseline resulted in only one statistically significant result (out of 45 tests).

For N03CC, despite reporting overwhelmingly positive results in the primary report, not all approaches produced a significant result. Considering only change from baseline, non-significant results for three of the five endpoints were noted when the total hip values were used, whereas the BMD results for spine and femoral neck were all significant. Percentage change from baseline indicated fewer statistically significant differences than for raw change from baseline (14 versus 39 each out of 45 tests). Consistent with N02C1, RawBMD, and TPerc provided the most statistically significant results (14/18 tests each).

Discussion

In this study, we examined the impact of five methods for calculating BMD endpoints on the likelihood of finding statistically significant differences between randomized treatment groups in patients with cancer. One of the main findings was that among the five endpoints, machine-based T score (RawT), sample-adjusted T score (TSamp), and reference population-adjusted T score (TRef) had 11, 8, and 7 statistically significant results out of 36 tests, respectively, whereas RawBMD and percentile T score (TPerc) produced 14 and 19 statistically significant results, respectively. Thus, in this case, the treatment comparisons had different conclusions depending upon the endpoint selected. Reference-based endpoints (RawT, TSamp, and TRef) aid in interpretation within the given study because values are reported in terms of standard deviations above or below a population mean. Transforming via sample statistics provided similar results as transforming via reference or machine-based norms. However, TSamp lacks the ability to be interpreted relative to an external reference population (or across trials). Specifically, baseline values appear similarly distributed between the two clinical trials in this report based on TSamp, whereas RawT and TRef properly show that patients on N03CC had lower baseline BMD relative to patients on N02C1. The new approach (TPerc) may be more sensitive to change and has the added benefit of being readily interpretable as a percentile.

In terms of anatomical site, spine, femoral neck, and total hip had 26, 20, and 13 statistically significant results across the two trials. The variability in results highlights that study conclusions could differ depending on the chosen anatomical site, consistent with recent studies that indicated using various measures of osteoporosis could cause different conclusions to be drawn [23].

Statistical testing based on change from baseline resulted in more statistically significant results relative to percentage change from baseline (39 versus 14). Percentage change from baseline endpoints generally suffer from lower statistical power relative to their respective change from baseline endpoint assuming normally distributed baseline and post-baseline scores with a range of correlations, do not protect from bias in the presence of baseline imbalances in the endpoint, and potentially violate normality assumptions of parametric statistical tests (e.g., *t* tests and ANCOVA) [31]. Percentage change from baseline endpoints is often employed nonetheless as endpoints likely due to the ease in interpretation of point estimates. While possible to construct non-normal data scenarios (i.e., data scenarios not investigated via simulation in the work of Vickers [31]) in which power may be similar or slightly improved by using percent change from baseline relative to change from baseline, results from the current report using observed clinical trial data revealed that statistical analysis based on change from baseline endpoints led to a greater number of statistically significant results compared to percent change from baseline. As suggested by Vickers [31], statistical testing based on change from baseline endpoints can be supplemented with percentage change point estimates (computed from mean baseline and follow-up scores) if percentage type endpoints are desired for reporting purposes.

We have focused our findings on selecting among endpoints, and less so on selecting among statistical tests (in this case between the non-parametric Wilcoxon rank-sum test and a

parametric two-sample *t* test or ANCOVA approach). The parametric two-sample *t* test will provide better power than the Wilcoxon rank-sum test under the assumption that data are normally distributed. An ANCOVA approach provides an additional gain in power over the two-sample *t* test if the selected covariate is correlated with the endpoint being compared. However, despite the loss of information in reducing data to their ranks, the Wilcoxon rank-sum test is only marginally less powerful than the two-sample *t* test under the assumption of normality, but often more powerful than the two-sample *t* test when the normality assumption fails [32] perhaps making the Wilcoxon rank-sum test the recommended choice to accommodate a wider range of possible data scenarios. All in all, selection of an appropriate test among the tests employed here as well as many other possible hypothesis testing procedures should always be based on the study design, expected distributional properties of the data, and ideally, on previous data exploration.

While we have identified that TPerc and analysis based on change from baseline are the most sensitive approaches, further study is needed to determine if this is desirable. We have provided direct empirical evidence that different answers regarding treatment efficacy are possible from different measurement approaches so caution in interpretation of study results is indicated. Researchers should consider calculating these measures in future trials as sensitivity analyses. Caution in conclusions based on the primary endpoint is indicated when sensitivity analyses show lack of consistency across endpoints. Researchers should also clearly report how endpoints were calculated in all publications due to the possible variability of results across endpoints.

This study also uncovered logistical hurdles that became lessons learned for future research. Analysis of data from BMD scanners at clinics without careful planning and forethought may lead to errors. In theory, it would be optimal to have the same scanner, same operator, same reference population, and the same site for providing BMD assessment for a given clinical trial, though this approach is impractical for most clinical studies, particularly multi-center studies such as the two clinical trials used in the current analysis. However, technical specification of the BMD assessment (e.g., clear definition of the anatomical sites to be assessed) in the protocol and well-defined data fields on case report forms can contribute to improved standardization. Multiple post hoc statistical techniques cannot overcome the inherent variability of the measurement process due to a lack of standardization.

Our study has limitations. First, as previously mentioned, our criterion for comparing endpoints and anatomical sites was based on the number of statistically significant results (i.e., sensitivity). This may not be an ideal strategy. However, our purpose in this analysis was to highlight that the choice of endpoint can impact study conclusions. A second limitation is that the endpoint and analyses in this report did not take into account variability by center, scanner, or other sources of known variation in BMD assessment. As these were both multi-center trials, it became apparent over the courses of these trials that BMD was not measured uniformly across sites. Possible sources of variation among centers included different machines, reference values, clinical report formats, and definitions of anatomical sites. Careful consideration of the data processes is vital to the science of the study. Standardization across centers may produce less variability across the endpoints examined in this report.

In conclusion, the five osteoporosis endpoint measures examined in this study are not necessarily the same and results differ across the anatomical sites tested. TPerc seems to be the most sensitive among the alternative statistical approaches. This new measure might be more practical than others because it is easily interpretable. Whether or not the seeming superior sensitivity is merely due to an increased type I error rate, however, is an open question in need of further study. Our ultimate recommendation is for future clinical trials to use multiple endpoints by way of a sensitivity analysis to ensure the credibility and veracity of overall study conclusions.

Acknowledgments

Financial support This work was supported in part by Public Health Service grants CA25224, CA37404, CA35431, CA35415, CA35103, and CA35269. The study was also supported, in part, by grants from the National Cancer Institute (CA31946) to the Alliance for Clinical Trials in Oncology (Monica M. Bertagnolli, M.D., Chair) and to the Alliance Statistics and Data Center (Daniel J. Sargent, Ph.D., CA33601).

References

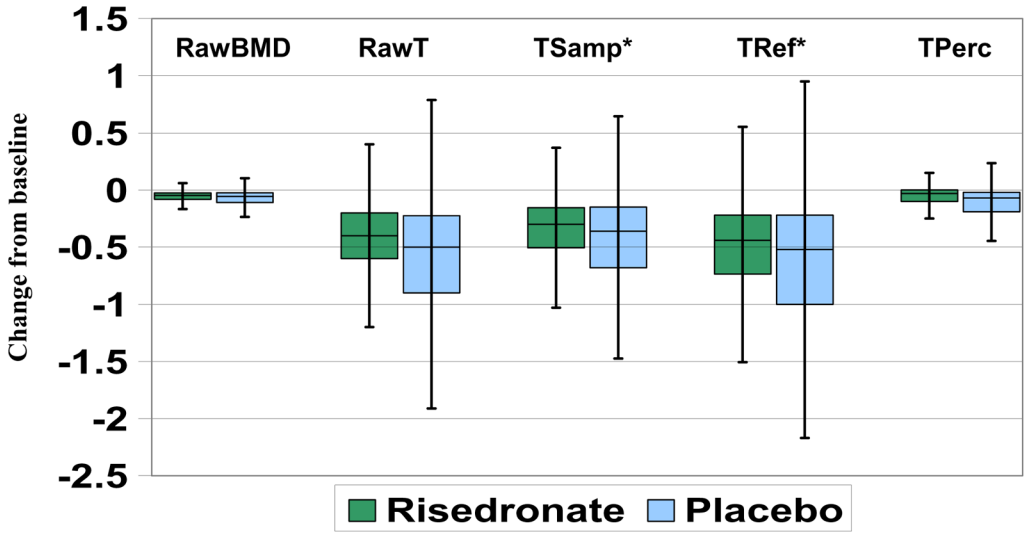
1. Johnell O. Advances in osteoporosis: better identification of risk factors can reduce morbidity and mortality. J Intern Med. 1996; 239(4):299–304. [PubMed: 8774383]
2. Ioannidis G, Papaioannou A, Hopman WM, Akhtar-Danesh N, Anastassiades T, Pickard L, Kennedy CC, Prior JC, Olszynski WP, Davison KS, Goltzman D, Thabane L, Gafni A, Papadimitropoulos EA, Brown JP, Josse RG, Hanley DA, Adachi JD. Relation between fractures and mortality: results from the Canadian Multicentre Osteoporosis Study. CMAJ. 2009; 181(5): 265–271.10.1503/cmaj.081720 [PubMed: 19654194]
3. Hasserius R, Karlsson MK, Nilsson BE, Redlund-Johnell I, Johnell O. Prevalent vertebral deformities predict increased mortality and increased fracture rate in both men and women: a 10-year population-based study of 598 individuals from the Swedish cohort in the European Vertebral Osteoporosis Study. Osteoporos Int. 2003; 14(1):61–68. [PubMed: 12577186]
4. Nochowitz B, Siegert S, Wasik M. An update on osteoporosis. Am J Ther. 2009; 16(5):437–445.10.1097/MJT.0b013e31818637de [PubMed: 19262365]
5. Valachis A, Polyzos NP, Georgoulias V, Mavroudis D, Mauri D. Lack of evidence for fracture prevention in early breast cancer bisphosphonate trials: a meta-analysis. Gynecol Oncol. 2010; 117(1):139–145. [PubMed: 20061004]
6. Leib ES, Lewiecki EM, Binkley N, Hamdy RC. Official positions of the International Society for Clinical Densitometry. J Clin Densitom. 2004; 7(1):1–6. [PubMed: 14742881]
7. Kanis JA. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report. WHO Study Group. Osteoporos Int. 1994; 4(6):368–381. [PubMed: 7696835]
8. Silberstein EB, Levin LL, Fernandez-Ulloa M, Gass ML, Hughes JH. Bone mineral density (BMD) assessment of central skeletal sites from peripheral BMD and ultrasonographic measurements: an improved solution employing age and weight in type 3 regression. J Clin Densitom. 2006; 9(3):323–328. [PubMed: 16931351]
9. Cole R, Larson J. The effect of measurement of the contralateral hip if the spine is not included in the bone mineral density analysis. J Clin Densitom. 2006; 9(2):210–216. [PubMed: 16785083]
10. Frost ML, Blake GM, Fogelman I. Can the WHO criteria for diagnosing osteoporosis be applied to calcaneal quantitative ultrasound? Osteoporos Int. 2000; 11(4):321–330. [PubMed: 10928222]
11. Abrahamsen B, Hansen TB, Jensen LB, Hermann AP, Eiken P. Site of osteodensitometry in perimenopausal women: correlation and limits of agreement between anatomic regions. J Bone Miner Res. 1997; 12(9):1471–1479. [PubMed: 9286764]
12. Arlot M, Meunier PJ, Boivin G, Haddock L, Tamayo J, Correa-Rotter R, Jasqui S, Donley DW, Dalsky GP, Martin JS, Eriksen EF. Differential effects of teriparatide and alendronate on bone

remodeling in postmenopausal women assessed by histomorphometric parameters. J Bone Miner Res. 2005; 20(7):1244–1253. [PubMed: 15940379]

13. Faulkner KG, von Stetten E, Miller P. Discordance in patient classification using T-scores. J Clin Densitom. 1999; 2(3):343–350. [PubMed: 10548828]
14. Grampp S, Genant HK, Mathur A, Lang P, Jergas M, Takada M, Glüer CC, Lu Y, Chavez M. Comparisons of noninvasive bone mineral measurements in assessing age-related loss, fracture discrimination, and diagnostic classification. J Bone Miner Res. 1997; 12(5):697–711. [PubMed: 9144335]
15. Greenspan SL, Bouxsein ML, Melton ME, Kolodny AH, Clair JH, Delucca PT, Stek M Jr, Faulkner KG, Orwoll ES. Precision and discriminatory ability of calcaneal bone assessment technologies. J Bone Miner Res. 1997; 12(8):1303–1313. [PubMed: 9258762]
16. Jergas M, Genant HK. Spinal and femoral DXA for the assessment of spinal osteoporosis. Calcif Tissue Int. 1997; 61(5):351–357. [PubMed: 9351874]
17. Clowes JA, Peel NF, Eastell R. Device-specific thresholds to diagnose osteoporosis at the proximal femur: an approach to interpreting peripheral bone measurements in clinical practice. Osteoporos Int. 2006; 17(9):1293–1302. [PubMed: 16810454]
18. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008; 61(2): 102–109. [PubMed: 18177782]
19. Chen Z, Maricic M, Lund P, Tesser J, Gluck O. How the new Hologic hip normal reference values affect the densitometric diagnosis of osteoporosis. Osteoporos Int. 1998; 8(5):423–427. [PubMed: 9850349]
20. Kanis JA. An update on the diagnosis of osteoporosis. Curr Rheumatol Rep. 2000; 2(1):62–66. [PubMed: 11123041]
21. Hines SL, Mincey BA, Sloan JA, Thomas SP, Chottiner E, Loprinzi CL, Carlson MD, Atherton PJ, Salim M, Perez EA. Phase III randomized, placebo-controlled, double-blind trial of risedronate for the prevention of bone loss in premenopausal women undergoing chemotherapy for primary breast cancer. J Clin Oncol. 2009; 27(7):1047–1053. [PubMed: 19075260]
22. Hines SL, Mincey B, Dentchev T, Sloan JA, Perez EA, Johnson DB, Schaefer PL, Alberts S, Liu H, Kahanic S, Mazurczak MA, Nikcevich DA, Loprinzi CL. Immediate versus delayed zoledronic acid for prevention of bone loss in postmenopausal women with breast cancer starting letrozole after tamoxifen-N03CC. Breast Cancer Res Treat. 2009; 117(3):603–609. [PubMed: 19214743]
23. Looker AC, Orwoll ES, Johnston CC Jr, Lindsay RL, Wahner HW, Dunn WL, Calvo MS, Harris TB, Heyse SP. Prevalence of low femoral bone density in older U.S. adults from NHANES III. J Bone Miner Res. 1997; 12(11):1761–1768. [PubMed: 9383679]
24. National Osteoporosis Foundation. [Accessed 25 June 2014] Clinician's guide to prevention and treatment of osteoporosis. 2014. http://nof.org/hcp/clinicians-guide
25. Perez EA, Josse RG, Pritchard KI, Ingle JN, Martino S, Findlay BP, Shenkier TN, Tozer RG, Palmer MJ, Shepherd LE, Liu S, Tu D, Goss PE. Effect of letrozole versus placebo on bone mineral density in women with primary breast cancer completing 5 or more years of adjuvant tamoxifen: a companion study to NCIC CTG MA.17. J Clin Oncol. 2006; 24(22):3629–3635. [PubMed: 16822845]
26. Reid IR, Brown JP, Burckhardt P, Horowitz Z, Richardson P, Trechsel U, Widmer A, Devogelaer JP, Kaufman JM, Jaeger P, Body JJ, Brandi ML, Broell J, Di Micco R, Genazzani AR, Felsenberg D, Happ J, Hooper MJ, Ittner J, Leb G, Mallmin H, Murray T, Ortolani S, Rubinacci A, Saaf M, Samsioe G, Verbruggen L, Meunier PJ. Intravenous zoledronic acid in postmenopausal women with low bone mineral density. N Engl J Med. 2002; 346(9):653–661. [PubMed: 11870242]
27. Schneider DL, Bettencourt R, Barrett-Connor E. Clinical utility of spine bone density in elderly women. J Clin Densitom. 2006; 9(3):255–260. [PubMed: 16931341]
28. Moore, DS. The basic practice of statistics. 3. W. H. Freeman; New York: 2004.
29. Altman, DG. Practical statistics for medical research. Chapman & Hall; London: 1991.
30. Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. Applied linear statistical models. 4. McGraw Hill; 1996.

31. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol. 2001; 1:6. [PubMed: 11459516]

32. Hodges JL, Lehmannn EL. The efficiency of some nonparametric competitors of the t-test. Ann Math Stat. 1956; 27:324–335.

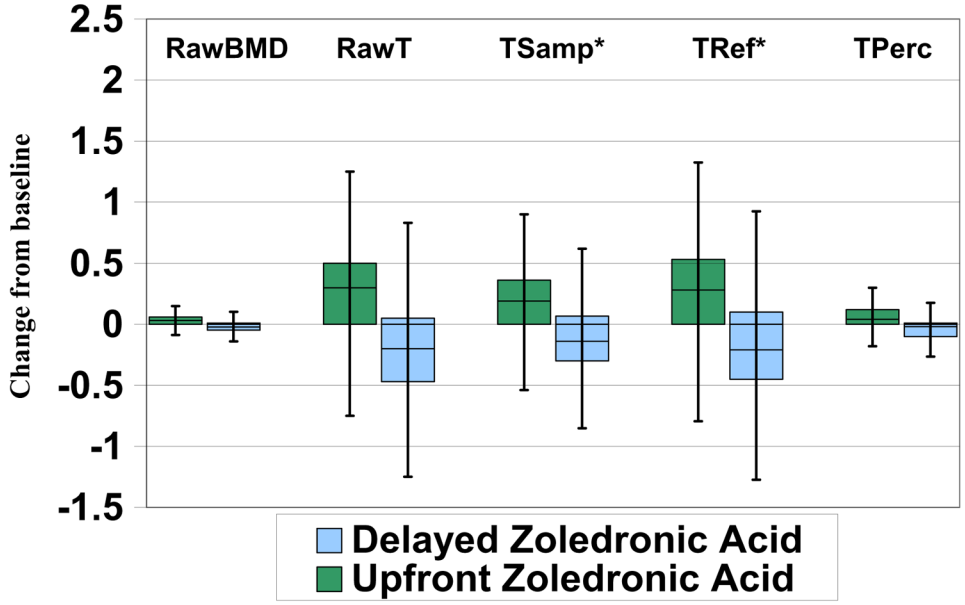


P-value					
T-test	0.43	0.15	0.43*	0.43*	0.04
Wilcoxon	0.32	0.13	0.32*	0.32*	0.02
ANCOVA	0.14	0.04	0.14*	0.14*	0.04

Fig. 1.

Boxplots showing the distribution of spine bone mineral density change from baseline endpoints by treatment arm for N02C1

* TSamp and TRef are linear transformations of RawBMD, resulting in identical results when comparing absolute (but not percentage) change from baseline between groups using a Wilcoxon rank-sum test, a two-sample t-test, or analysis of covariance (ANCOVA).



P-value					
T-test	<0.001	<0.001	<0.001*	<0.001*	<0.001
Wilcoxon	<0.001	<0.001	<0.001*	<0.001*	<0.001
ANCOVA	<0.001	<0.001	<0.001*	<0.001*	<0.001

Fig. 2.

Boxplots showing the distribution of spine bone mineral density change from baseline endpoints by treatment arm for N03CC

* TSamp and TRef are linear transformations of RawBMD, resulting in identical results when comparing absolute (but not percentage) change from baseline between groups using a Wilcoxon rank-sum test, a two-sample t-test, or analysis of covariance (ANCOVA).

Table 1

Mean (standard deviation) for bone mineral density parameters from different machines used for the calculation of T scores

Bone mineral density parameter	Local reference population	Manufacturers' values
Hologic Sahara speed of sound (m/s)	1560.7 (25.1)	not available
Hologic UBA575+ speed of sound (m/s)	1507.5 (6.6)	not available
Osteometer DTUone speed of sound (m/s)	1553.5 (8.4)	not available
Hologic Sahara heel (g/m^2)	0.561 (0.10)	0.537 (0.08)
Dual x-ray absorptiometry lumbar spine (g/m^2)	1.068 (0.12)	1.047 (0.11)
Dual x-ray absorptiometry femoral hip (g/m^2)	0.892 (0.10)	0.895 (0.10)
Dual x-ray absorptiometry total hip (g/m^2)	0.988 (0.10)	0.975 (0.12)

From [10]

Table 2

Mean and standard deviation (SD) for five spine bone mineral density endpoints at baseline in two North Central Cancer Treatment Group studies (N02C1 and N03CC)

Endpoint	N02C1 *N*=204		N03CC *N*=505	
	Mean	SD	Mean	SD
RawBMD	1.21	0.16	1.10	0.16
RawT	0.46	1.19	−0.17	1.18
TSamp	0.00	1.00	0.00	1.00
TRef	1.51	1.44	0.50	1.49
TPerc	0.80	0.26	0.59	0.32

Table 3

Number of significant statistical tests obtained from three different statistical procedures for each combination of anatomical site, endpoint, and clinical trial (N02C1 and N03CC)

Endpoint	N02C1							N03CC							Total
	Change from Baseline			% Change from Baseline			Total	Change from Baseline			% Change from Baseline			Total	
	Spine	Fem. neck	Total hip	Spine	Fem. neck	Total hip		Spine	Fem. neck	Total hip	Spine	Fem. neck	Total hip		
RawBMD	0	0	0	0	0	0	*0*	3	3	1	3	3	1	14	*14*
RawT	1	0	0	0	0	0	*1*	3	3	3	1	0	0	10	*11*
TSamp	0*	0*	0*	0	0	0	*0*	3*	3*	1*	1	0	0	8	*8*
TRef	0*	0*	0*	0	0	0	*0*	3*	3*	1*	0	0	0	7	*7*
TPerc	3	1	0	1	0	0	*5*	3	3	3	1	1	3	14	*19*
Total	*4/15*	*1/15*	*0/15*	*1/15*	*0/15*	*0/15*	*6/90*	*15/15*	*15/15*	*9/15*	*6/15*	*4/15*	*4/15*	*53/90*	*59/180*

* TSamp and TRef are linear transformations of RawBMD, resulting in identical results when comparing absolute (but not percentage) change from baseline between groups using a Wilcoxon rank-sum test, a two-sample *t* test, or analysis of covariance (ANCOVA)