TITLE:

THE POTENTIAL VALUE OF MONITORING BONE TURNOVER MARKERS AMONG WOMEN ON ALENDRONATE

AUTHORS:

Katy J.L. Bell, Andrew Hayen, Les Irwig, Marc C. Hochberg, Kristine E. Ensrud, Steve Cummings, Douglas C. Bauer

ABSTRACT (240 words)

Biochemical markers of bone turnover have been proposed to monitor the response to bisphosphonate therapy for osteoporosis, but this requires true between-person differences in the response to therapy.. Using mixed models we analysed 3 annual measurements of two markers (bone alkaline phosphatise (BAP) and cross linked N-telopeptide of type I collagen (NTX)) from the Fracture Intervention Trial. We compared marker variation among women allocated to alendronate with that among women allocated to placebo to estimate how much variation was due to true between-person differences in response to treatment, and how much was due to random within-person fluctuations unrelated to treatment. For both markers we found that the mean effect of treatment differed by the baseline level of marker. After allowing for this and other effects, we found large true between-person differences in response to treatment for both markers, with a coefficient of variation for NTX of 25.1% and for BAP of 21.2%. However random within-person fluctuation was even larger, with a coefficient of variation for change in NTX of 42.5% and for change in BAP of 25.8%. Although repeated measurements have the potential to reduce within person variability, even triplicate baseline marker measurements resulted in an averaged value that was only within 30% of the true value with 95% certainty. In summary, although bone turn-over markers appear promising for monitoring between person differences in response to treatment, their use in clinical practice is currently limited by large random within-person variation.

WORD COUNT

Text: 3899

INTRODUCTION

Alendronate and other N-containing bisphosphonates are effective therapies(1-2) that are widely used for prevention of fragility fractures in postmenopausal women with osteoporosis. After starting a patient on alendronate, clinicians often monitor the response to make sure treatment is working as expected. Although monitoring bone mineral density is widely used to check treatment response, and routine monitoring in the first two years of treatment is recommended by the US National Osteoporosis Foundation and the American Association of Clinical Endocrinologists, recent evidence questions the value of this practice(3-6). Bone turnover markers (BTMs) have been proposed as an alternative way of monitoring treatment response(7), but the value of this is currently unknown.

Bone turnover may be estimated by markers of bone formation (e.g. bone-specific alkaline phosphatise (BAP) and N-terminal propeptide of type I collagen (PINP)) and bone resorption (e.g. cross-linked N-telopeptide of type I collagen (NTX) and cross-linked C-telopeptide of type I collagen (CTX)). Meta-analytic data(8) and individual trial data(7, 9) demonstrate associations between treatment effects on bone turnover markers and treatment effects on fracture risk. Bone turnover markers have the advantages of relative ease and low cost for a single measurement, as well as a rapid response to treatment, especially when compared to bone mineral density measurement. Their main disadvantage is the large within person random variation which results from both pre-analytic and analytic sources(10).

Guidelines for treatment of post-menopausal osteoporosis differ in their recommendations for the routine monitoring bone turnover markers after starting bisphosphonates. The American Association of Clinical Endocrinologists (U.S.)(11) and the National Osteoporosis Foundation (U.S.)(12) recommend that

markers of bone turnover may be useful for assessing therapeutic responses to antiresorptive agents including bisphosphonates. A Consensus Development Panel sponored by the National Institute of Health (U.S.)(13) recommends against changing therapy because of an adverse trend in bone turnover. The National Osteoporosis Guidelines Group (U.K.)(14) do not make recommendations regarding the monitoring of bone turnover markers and the Osteoporosis Society of Canada (Canada)(15) recommends that although biomarkers may be of value in predicting and monitoring response to potent antiresorptive therapy in clinical trials, they should not yet be used for routine clinical management.

The objective of the present analysis was to assess the potential value of monitoring bone turnover markers for determining an individual's response to alendronate therapy. We achieved this by comparing the variation in two biochemical markers, serum BAP (a marker of bone formation) and urine NTX (a marker of bone resorption) among subjects randomized to alendronate or placebo in the Fracture Intervention Trial (FIT). This allowed the overall variation in response to be separated into that due to true between-person variation in response and that due to random background within-person variation. Response monitoring may be clinically useful when true between-person variation in response is large and random background within-person variation in the marker is small.

METHODS

Study Design and Population

We analysed data from the Fracture Intervention Trial (FIT), a randomized trial that evaluated the effects of alendronate compared to placebo in post-menopausal women with low bone mineral density (≤ 0.68 g/cm^2 at baseline)(16). The trial had two arms: the vertebral fracture arm, which included 2027 women who had vertebral fractures identified on radiographs at baseline(1), and the clinical fracture arm, which included 4432 women without baseline vertebral fractures(2). Patients for both arms were recruited between May 1992 and May 1993. A random sample of women in FIT (n= 1304 of 6459, 20.2%) had

serial measurements of biochemical markers of bone turnover, comprising 392 women from the Vertebral Fracture Arm (1) and 912 women from the Clinical Fracture Arm (2).

Subjects were randomly allocated to daily alendronate or placebo. The dose of alendronate was initially 5 mg/day for two years but was increased to 10mg/day at the second annual visit because other trials suggested that 10mg/day had greater effects on bone mineral density. Women in each treatment group who had dietary calcium intakes <1000 mg/day at baseline (82% of participants) were asked to take a daily supplement (OsCal0 containing 500mg of elemental calcium and 250 IU of vitamin D).

Monitoring Measurements

Biochemical markers of bone turnover

Fasting serum and a spot early morning urine samples were collected at baseline and yearly intervals following randomisation and stored at -70C for later analysis. Serum was analysed for bone specific alkaline phosphatase (BAP, assay from Tandem, Hybritech, Inc., San Diego, CA). This immuno radiometric assay (IRMA) uses two monoclonal antibodies directed against the human bone isoenzyme and BAP purified from human SAOS-2 osteosarcoma cells as a standard and has a 16% cross-reactivity with the circulating liver isoenzyme(17). Urine was analysed for N-teleopeptide of type I collagen (NTX, assay from Osteomark®, Ostex International, Inc., Seattle, WA), with a correction applied for the woman's level of urinary creatinine. This enzyme-linked immunoassay (ELISA) uses a monoclonal antibody directed against the N-telopeptide-to-helix intermolecular cross-linking domain of type I collagen isolated from human urine(18). The reported intra- and inter-assay coefficients of variation are both <10% for BAP(17) and urinary NTX(19). All measurements were performed by a commercial laboratory (Nichols Institute) contemporaneously for alendronate and placebo groups in annual batches without knowledge of treatment assignment. Yearly measurements were available from baseline to three years of follow up for women in both the Vertebral Fracture Arm and the Clinical Fracture Arm.

Statistical Analysis

For the preliminary analysis we applied a simple comparison method(20) to summary data for each marker on the log-scale with both arms of FIT combined whereby we compared marker variation in women allocated to alendronate with that in women allocated to the placebo. We subtracted the mean and variance of change in the placebo group from the mean and variance of change in the alendronate group and back transformed the results to the natural scale.

For the main analysis, we used a type of statistical model known as the 'mixed model'(21-22). In these models we again included data from both arms of FIT, but adjusted for the trial arm of a particular participant by including terms to represent trial arm and a trial arm × treatment interaction. Mixed models allowed explicit modelling of the within and between person variation in the bone turnover markers, while also taking into account the correlation between measurements taken on the same individual. These models also allowed for some individuals to have missing follow up measurements by using whatever data is available for each individual. We fitted a series of mixed models using bone turnover measurements over three years, with measurements from years 1-3 as the outcome and baseline measurement as a predictor. A natural log transformation was applied to both outcomes in order to normalise the distributions of residuals. After transformation, the assumptions of normality, linearity and homoscedascity were found to hold at both the individual level and the residual level.

Likelihood ratio tests were used to compare models where treatment effect on bone turn-over was the same for everyone (treatment had a fixed effect) with models where treatment effect on bone turn-over differed between individuals (treatment had random effects) in order to assess statistical evidence for variation in treatment response. For fixed effects we estimated the mean effect applicable to all patients, whereas for random effects we estimated the mean and standard deviation of effects across patients. For random effects we also estimated a 95% distribution of treatment effects, an interval in which the treatment effect for 95% of the study population should lie (this was expressed as a % increase or decrease on the natural scale after back-transforming the mean +/-1.96 x standard deviation, and

subtracting 1). We used the residuals to calculate the coefficient of variation for within person variation of one measurement (estimated by back-transforming the standard deviation of within person variation and subtracting 1(23), CV_W) and of change between two measurements (estimated as $\sqrt{}(2xCV_W^2)$). We compared this to the coefficient of variation for between person variation in treatment effects (estimated by back-transforming the standard deviation of treatment effects and subtracting 1). In this way we estimated the mean treatment effect, the true between person variation in treatment effects and the background random within person variation for change in each bone turnover marker. We also compared the ratio of variances for between person variation in treatment effects: within person variation in change on the log scale as an indication of signal: noise ratio (For more detailed explanations of this type of statistical analysis, see (3, 24)).

As the primary analysis we fitted BAP and NTX as separate outcomes (univariate response models). We also fitted models that considered both markers simultaneously while allowing for their correlation (bivariate response models). Analysis was done using MLwiN with models fitted using iterative generalised least squares (Centre for Multilvel Modelling, University of Bristol, U.K.)

The main analysis described above was by intention to treat, and all subjects were included regardless of compliance with study medication or fracture outcomes. We also performed sensitivity analyses where we limited data to either women who had been at least 75% adherent throughout the trial (as estimated by pill-count) or to women who had complete baseline and follow up data available for the relevant marker.

Clinicians often like to establish baseline levels of a marker against which to judge change after treatment. We used the mixed model estimates of within person variation to estimate the number of measurements needed to be 95% certain that an observed baseline level of marker was within 10% or 20% of the woman's true baseline level, as well as the level of uncertainty that would result with single, duplicate or triplicate measurement. We then used the estimates of treatment effects and within person

variation to calculate the number of measurements needed to be 95% certain that an apparent reduction in bone turn over greater than 20%, 30% and 40% reflects the true reduction due to treatment.

A more detailed explanation of how mixed models were used and the calculations used for estimating the number of measurements needed are given in the Appendix.

Role of the funding source

This analysis was funded by the Australian National Health and Medical Research Council (Program Grant No 402764). The Fracture Intervention Trial was sponsored by Merck Research Laboratories. The sponsors did not influence the design and conduct of the study; collection, management, analysis and interpretation of data; preparation and approval of manuscript.

RESULTS

Summary statistics for levels and change from baseline and 1 year are presented in Table 1. There was a mean % decrease in the placebo group, likely to reflect in part the provision of calcium and vitamin D supplementation provided to women with calcium intake <1000 mg. Baseline NTX levels were similar for the Vertebral Fracture Arm (63.6 nmol/ mmol Cr) and the Clinical Fracture arm (63.3 nmol/ mmol Cr, p=0.91); baseline BAP levels were slightly higher for the Vertebral Fracture Arm (13.4 ng/ml) than for the Clinical Fracture Arm (11.6 ng/ml, p<0.001). The wide spread of % change in markers in the placebo group (represented by the 2.5 and 97.5 percentiles), reflects the substantial background random within person variation that exists for these bone turn over markers. The larger mean % decreases in the alendronate group compared with the placebo group reflects the average effects of treatment. The larger spread of % change in markers in the alendronate group compared with the placebo group (2.5 to 97.5 percentiles) reflects the between-person variation in treatment effects. Applying the simple comparison method(20), we estimated the true mean treatment effect on NTX was 50%↓ and the 95% distribution of

true treatment effects was 75%↓ to 1%↑.The estimated true mean treatment effect on BAP was 32%↓ and the 95% distribution of true treatment effects was 58%↓ to 10%↑.

The mixed models used measurements from both treatment groups from baseline to year 3 of follow up. The number of measurements available for each marker, stratified by treatment group is summarised in the Appendix Table. A summary of the results of the mixed model analysis follows.

Treatment effects and variability of NTX

Treatment effects on NTX are summarised in Table 2, upper rows. The mean treatment effect on NTX was substantial for all subgroups and slightly greater for women in the Vertebral Fracture Arm (mean decrease in NTX levels at 1 year = 56.1%) compared with women in the Clinical Fracture Arm (mean decrease in NTX levels at 1 year = 52.1%), ($p=0.010$) There was a small increase in the mean treatment effect with time for both trial arms, likely to be due at least in part to the increase in dose after the first 2 years ($p<0.001$). The mean treatment effect was also modified by baseline NTX levels, with greater reduction for women with higher baseline levels ($p<0.001$,).

After allowing for the modification of mean treatment effects outlined above, and for other significant predictors, there was very strong evidence of large between person variation in treatment effects ($p<0.001$, coefficient of variation 25.1%). The variation was more noticeable for women with lower baseline levels of NTX. However none of the 95% distributions of treatment effects crossed zero, even for women with the lowest baseline level of NTX (8 nmol/L, 2.5th percentile of treatment effects=9.9% decrease for women in the Clinical Fracture Arm). The coefficient of variation for background within person variation was very large at 30.0% for one measurement and 42.5% for change between two measurements. The variance ratio for between person variation in treatment effects: within person variation in change was 0.31 (variation in treatment effects was just under one third as large as within person variation).

Treatment effects and variability of BAP

Treatment effects on BAP are summarised in Table 2, lower rows. The mean treatment effect on BAP was much smaller than for NTX and differed by a larger amount between trial arms. The mean treatment effect on BAP was smaller for women in the Vertebral Fracture Arm (mean decrease in BAP levels at 1 year = 13.8%) compared with women in the Clinical Fracture Arm (mean decrease in BAP levels at 1 year = 24.9%, p value for difference <0.001). There was again a small increase in the mean treatment effect with time for both trial arms, likely to be due at least in part to the increase in dose after the first 2 years (p<0.001). The mean treatment effect was also modified by baseline BAP levels, with greater reduction for women with higher baseline levels (p<0.001).

After allowing for the modification of mean treatment effects outlined above, and for other significant predictors, there was very strong evidence of large between person variation in treatment effects (p<0.001, coefficient of variation 21.2%). This variation was especially marked for women with lower baseline BAP levels, but treatment effects crossed 0% change for all sub-groups of women except those with higher than average baseline levels who did not have a baseline vertebral fracture (this subgroup had the largest mean effect of treatment – see Table 2, bottom row). Some of these women had a substantial net increase in BAP level after starting treatment (we estimate a true increase in BAP of up to 41.9% for 2.5% of women with a vertebral fracture and baseline levels of BAP around 2 ng/ml).

The coefficient of variation for background within person variation was 18.2% for one measurement and 25.8% for change between two measurements. The variance ratio for between person variation in treatment effects: within person variation in change was 0.58 (variation in treatment effects was just over half as large as within person variation).

Additional analyses

The bivariate model (where BAP and NTX were considered simultaneously) yielded similar results, with a very high correlation between the effects of treatment on BAP and NTX at 1 year (0.99).

Limiting analysis to women who were adherent over the course of the trial (n=1032) revealed similar mean treatment effects (slightly greater overall, though slightly smaller for women with lowest baseline levels), somewhat smaller between-person variation in treatment effects and similar levels of within-person variation. There was still an increase in BAP as a result of treatment for some women, although the number affected was less than in the intention to treat analysis. Limiting analysis to women for whom we had complete marker data (n=775 for NTX and n=1032 for BAP), yielded results more similar to the intention to treat analysis (data not shown).

Effect of repeated measurements

To provide an estimate of the effects of large within person variability on clinical utility we estimated the effect of repeated measurements on marker accuracy. We limited calculations of the number of measurements needed to be certain of a given response to the urinary NTX marker because the distribution of treatment effects on BAP crossed 0% change for a large number of individuals.

We estimated that the average of 27 measurements are needed to be 95% certain that an observed baseline level of marker was within 10% of the woman's true baseline level and 7 measurements are needed to be 95% certain that an observed level was within 20% of the woman's true baseline level. If only a single measurement was available, the observed baseline level of marker would be within 55% of the woman's true baseline level with 95% certainty. If duplicate or triplicate measurements were available, the average of these measurements would be within 38% and 31% of the woman's true baseline level respectively, with 95% certainty. With 80% certainty, the corresponding values would be within 34%, 24% and 20% for single, duplicate and triplicate measurements.

For a woman without an existing vertebral fracture and a true baseline level of urinary NTX of 8 nmol/mmol Cr , we estimate that one measurement after starting treatment is needed to be 95% certain that an apparent decrease of more than 20% indicates a true decrease of more than 20%. Ten measurements after starting treatment are needed to be 95% certain that an apparent decrease of more than 40% indicates a true decrease of more than 40% and 56 measurements are needed to be 95% certain that an apparent decrease of more than 60% indicates a true decrease of more than 60%. Conversely, for a woman with an existing vertebral fracture and a true baseline level of urinary NTX of 158 nmol/mmol Cr, we estimate that only one measurement after starting treatment is needed to be 95% certain that apparent decreases of more than 20%, 40% and 60% indicate true decreases of more than 20%, 40% and 60% respectively.

DISCUSSION

In agreement with others' findings, we found that the mean effect of treatment was a decrease in markers of both bone resorption and bone formation(25). The mean effect of treatment appeared to differ depending upon whether or not a woman had a vertebral fracture at baseline (differences were more marked for BAP, with less treatment effect for women with a baseline fracture)and the baseline level of the marker (increasing treatment effect with increasing baseline marker level for both NTX and BAP).

After allowing for background variation both between and within individuals that is unrelated to treatment, we found very strong evidence of large variation between individuals in the true effects of alendronate on two markers of bone resorption and formation, particularly for women with low baseline marker levels. This is in contrast to our previous report on bone mineral density(3) from the same study where we found small variation in treatment effects between individuals which was probably not clinically relevant. Thus, as opposed to serial measurements of bone density, serial measurements of biochemical markers of bone turnover have the potential to provide useful information about individual response to bisphosphonate therapy.

Despite substantial between-person variation in treatment effects, this was muted by the large background random within-person variation in both markers, with CVs for change in BAP and NTX of 25.8% and 42.5 %. These figures are based on within-person SDs for a single measurement occasion of BAP and NTX of 18.2% and 30%, respectively, similar to estimates reported previously for these two markers (30-33). This large within person variation means that the number of measurements of NTX needed to be certain of a woman's baseline level of the marker is prohibitively large (7 measurements are needed for the observed values to be within 20% of the true level with 95% certainty; if 3 measurements are used then observed levels are only within 31% of the true level with 95% certainty). This means that monitoring change in the marker is challenging, even though under certain circumstances, relatively few (sometimes only one) repeated measurements are needed after treatment to estimate the true treatment effect once we are certain of the baseline level. We are unable to directly compare within person variation in the bone turn over markers (CV) with that in bone mineral density (SD) published in a previous report(3) as variation was not proportional to the mean level with the latter (and consequently CV will not stay constant as the mean level changes). However a comparison of the signal (between-person variance in treatment effects) to noise (within person variance) ratios for bone turn over markers (0.58 for BAP and 0.31 for NTX) and BMD (0.08, calculated from variances reported in Appendix 2 of (3)) suggests that markers show more potential for monitoring response to treatment.

The between person variation in treatment effects on BAP crossed 0% change in most of the subgroups examined. Overall treatment resulted in an increase in this marker of bone formation in approximately 16.7% of women. This finding was unexpected: despite reports that bisphosphonates may stimulate osteoblast formation(26-27), there is no histomorphometric evidence of an increase in bone formation with treatment(28) and in general bisphosphonates are thought to be anti-resorptive and not anabolic(29). Other studies, similar to the present one, are needed to support or refute this finding. Although there was variation in the % decrease in NTX there were no increase in levels in any sub-group of women, which is consistent with the primary anti-resorptive effect of alendronate and other bisphosphonates.

Strengths of our study include the randomised, placebo controlled nature of the data used and the relatively large number of women and measurement occasions that form the dataset. The very high rate of adherence in FIT means that the variation in treatment effects is likely to reflect differences between women in the actual effects of treatment. Lower adherence rates in clinical populations may increase this variation further. However the reservations regarding the clinical use of bone turn over markers to monitor response because of large within-person variation also apply to their use to monitor adherence, and direct questioning of the patient is probably a better method of determining this(34-35). Our study is limited by markers obtained during the study which did not included newer assays such as PINP. Further, only annual measurements were available.

The measurement of bone turn over markers was done in annual batches which may have allowed for mean drift over time and increased within-person variation compared to if all assays were done in the same batch. This issue applies equally to alendronate and placebo groups, so the estimates of treatment effects should be un-biased. It also reflects what happens in actual clinical practice, and if anything within-person variation is likely to be under-estimated in the trial setting because of minimisation of pre-analytic sources of variation (through standardisation of timing of specimen collection for example). Our estimates are therefore likely to be best case scenario estimates of the value of response monitoring using bone turn over markers in clinical populations.

For the first two years of the study, a lower dose of alendronate (5 mg) was used than is usually taken in clinical practice. The mean increase in treatment effect with time is likely to be due in large part to the increase in dose at 3 years (to 10 mg). There was no change in the variation due to treatment with time, suggesting the increase in dose did not affect variation in treatment response. Our estimates of between person variation and within person variation are likely to be similar to those if the higher dose was used for all 3 years.

In conclusion, we found very strong evidence of substantial between-person variation in the effects of alendronate on both a marker of bone resorption and a marker of bone formation. Thus, bone turnover over could represent a potentially useful method to monitor bisphosphonate treatment. However given the large within person variation that currently exists for these markers, they are likely to be of limited value for monitoring in clinical practice.

REFERENCES

1. Black D, Cummings S, Karpf D, Cauley J, Thompson D, Nevitt M, et al. Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Lancet. 1996;348(9041):1535-41.
2. Cummings S, Black D, Thompson D, Applegate W, Barrett-Connor E, Musliner T, et al. Effect of alendronate on risk of fracture in women with low bone density but wihtout vertebral fractures. JAMA. 1998;280(24):2077-82.
3. Bell KJL, Hayen A, Macaskill P, Irwig L, Craig JC, Ensrud KEE, et al. Value of routine monitoring of bone mineral density after starting bisphosphonate treatment: secondary analysis of trial data. BMJ 2009;338:b2266.
4. Cummings S, Palermo L, Browner W, Marcus R, Wallace R, Pearson J, et al. Monitoring Osteoporosis Therapy with Bone Densitometry: Misleading Changes and Regression to the Mean. JAMA. 2000;283:1318-21.
5. Pavlic M, Brand B, Cummings S. Estimating probability of non-response to treatment using mixture distributions. Statistics in Medicine 2001;20:1739-53.
6. Chapurlat RD, Palermo L, Ramsay P, Cummings S. Risk of fracture among women who lose bone density during treatment with alendronate. The Fracture Intervention Trial Osteoporosis International 2005;16:842-8.
7. Bauer D, Black DM, Garnero P, Hochberg M, Ott S, Orloff J, et al. Change in bone turnover and hip, non-spine, and vertebral fracture in alendronate-treated Women: The Fracture Intervention Trial. Journal of Bone and Mineral Research. 2004;19:1250-8.
8. Hochberg MC, Greenspan S, Wasnich RD, Miller P, Thompson DE, Ross PD. Changes in bone density and turnover explain the reductions in incidence of nonvertebral fractures that occur during treatment with antiresorptive agents. The Journal of Clinical Endocrinology & Metabolism 87(4):1586–92.
9. Bjarnason NH, Sarkar S, Duong T, Mitlak B, Delmas PD, Christiansen C. Six and twelve month changes in bone turnover are related to reduction in vertebral fracture risk during 3 years of raloxifene treatment in postmenopausal osteoporosis. Osteoporos Int Osteoporos Int (2001) 12:922–930;2:922–30.
10. Compston J. Monitoring osteoporosis treatment. Best Practice & Research Clinical Rheumatology 2009;23:781-8.
11. AACE Osteoporosis Task Force. American association of clinical endocrinologists medical guidelines for clinical practice for the prevention and treatment of postmenopausal osteoporosis: 2001 Edition, with selected updates for 2003 Endocrine Practice.9(6):544-64.
12. National Osteoporosis Foundation. Clinician's guide to prevention and treatment of osteoporosis 2008: Available from: http://www.nof.org.
13. NIH Consensus Development Panel on Osteoporosis Prevention Diagnosis and Therapy. Osteoporosis Prevention, Diagnosis, and Therapy. JAMA. 2001;285(6):785-95.
14. The National Osteoporosis Guidelines Group (U.K.). Osteoporosis clinical guideline for prevention and treatment. 2008.

15. Brown JP, Josse RG, Scientific Advisory Council of the Osteoporosis Society of Canada. 2002 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada. Canadian Medical Association. 2002;167(10 supplement):S1-S34.
16. Black DM, Nevitt MC, Cauley J, Karpf D, Cummings SR. Design of the Fracture Intervention Trial. Osteoporosis Internation. 1993;3 (Supplement 3):S29-S39.
17. Garnero P, Delmas PD. Assessment of the serum levels of bone alkaline phosphatase with a new immunoradiometric assay in patients with metabolic bone disease. Journal of Clinical Endocrinology and Metabolism. 1993;77:1046-53.
18. Hanson DA, Weiss MAE, Bollen AM, Maslan SL, Singer FR, Eyre DR. A specific immunoassay for monitoring human bone resorption: quantitation of type I collagen cross-linked N-telopeptides in urine. Journal of Bone and Mineral Research. 1992;7:1251–8.
19. Garnero P, Sornay-Rendu E, Duboeuf F, Delmas PD. Markers of bone turnover predict postmenopausal forearm bone loss over 4 years: The OFELY study. Journal of Bone and Mineral Research. 1999;14(9):1614-21.
20. Bell KJL, Irwig L, Craig JC, Macaskill P. Use of randomized trials to decide when to monitor response to new treatment. BMJ. 2008;336:361-5.
21. Goldstein H, Browne W, Rabash J. Multilevel modelling of medical data. Statistics in medicine. 2002;21:3291-315.
22. Singer J, Willet J. Exploring Longitudinal Data on change. Applied Longitudinal Data Analysis. New York: Oxford University Press, Inc.; 2003. p. 42-4.
23. Bland JM, Altman DG. Measurement error proportional to the mean. BMJ. 1996;313:106.
24. Bell KJL, Hayen A, Macaskill P, Craig JC, Neal BC, Irwig L. Mixed models showed no need for initial response monitoring after starting anti-hypertensive therapy. Journal of Clinical Epidemiology. 2009;62:650-9.
25. Bauer DC, Black DM, Garnero P, Hochberg M, Ott S, Orloff J, et al. Change in bone turnover and hip, non-spine, and vertebral fracture in alendronate-treated women: the Fracture Intervention Trial. Journal of Bone and Mineral Research. 2004;19:1250 –8.
26. Giuliani N, Pedrazzoni M, Negri G, Passeri G, Impicciatore M, Girasole G. Bisphosphonates stimulate formation of osteoblast precursors and mineralized nodules in murine and human bone marrow cultures in vitro and promote early
osteoblastogenesis in young and aged mice In vivo. Bone. 1998;22 455–61.
27. Viereck V, Emons G, Lauck V, Frosch KH, Blaschke S, Grundker C, et al. Bisphosphonates pamidronate and zoledronic acid stimulate osteoprotegerin production by primary human osteoblasts. Biochemical and Biophysical Research Communications 2002;291:680–6.
28. Chavassieux PM, Arlot ME, Reda C, Wei L, Yates AJ, Meunier PJ. Histomorphometric assessment of the long-term effects of alendronate on bone quality and remodeling in patients with osteoporosis. Journal of Clinical Investigation. 1997;100:1475–80.
29. Chapurlat RD, Delmas PD. Drug Insight: bisphosphonates for postmenopausal osteoporosis. Nature Clinical Practice Endocrinology & Metabolism 2006;doi:10.1038/ncpendmet0121.
30. Fink E, Cormier C, Steinmetz P, Kindermans C, Le Bouc Y, Souberbielle J-C. Differences in the capacity of several biochemical bone markers to assess high bone turnover in early menopause ad response to alendronate therapy. Osteoporos International. 2000;11:295–303.
31. Scarianoa JK, Garryb PJ, Montoyac GD, Wilson JM. Critical differences in the serial measurement of three biochemical markers of bone turnover in the sera of pre- and postmenopausal women. Clinical Biochemistry 2001;34:639–44.
32. Gertz BJ, Shao P, Hanson DA, Quan H, Harris ST, Genant HK, et al. Monitoring bone resorption in early postmenopausal women by an immunoassay for cross-linked collagen peptides in urine. Journal of Bone and Mineral Research. 1994;9:135-42.
33. Eastell R, Mallinak N, Weiss S, Ettinger M, Pettinger M, Cain D, et al. Biological variability of serum and urinary N-telopeptides of type I collagen in postmenopausal women. Journal of Bone and Mineral Research. 2000;15:594–8.

34. Haynes RB, Taylor DW, Sackett DL, Gibson ES, Bernholz CD, Mukherjee J. Can simple clinical measurements detect patient compliance Hypertension 1980;2:757-64.
35. Haynes RB, Sackett DL, Taylor DW. How to detect and manage low patient compliance in chronic illness. Geriatrics 1980;35.(1):91-7.

Table 1: Observed changes in Alendronate and Placebo groups at one year

Therapy		Baseline Level	One year Level	% Change from baseline*
NTX(nmol/ mmol Cr)				
Alendronate	Number			
		465	427	
	Mean	64.6	25.0	63%↓
	Median	57.0	20.1	
	2.5 to 97.5			88%↓ to 23%↑
	Percentile	21.7 to 145.4	8.0 to 73.5	
Placebo				
	Number	466	432	
	Mean	62.3	47.4	26%↓
	Median	57.7	42.7	
	2.5 to 97.5			66%↓ to 83%↑
	Percentile	22.5 to 126.3	16.3 to 114.0	
BAP (ng/ml)				
Alendronate	Number	626	594	
	Mean	12.1	7.3	41%↓
	Median	11.4	6.8	
	2.5 to 97.5			
	Percentile	5.2 to 22.9	2.8 to 15.3	72%↓ to 14%↑
Placebo				
	Number	656	630	
	Mean	12.1	10.5	14%↓
	Median	11.5	10.1	
	2.5 to 97.5			
	Percentile	5.4 to 22.2	4.2 to 19.8	53%↓ to 48%↑

*% change is calculated by back transforming change on log-scale

Table 2: True treatment effects after one year of treatment estimated from mixed models*

	Baseline level (ng/ml) †	Treatment effects on marker (Vertebral Fracture Arm)		Treatment effects on marker (Clinical Fracture Arm)		Between person variation in treatment effects: within person variation		
		Mean	95% distribution of effects	Mean	95% distribution of effects	Between person CV for treatment effects	Within person CV for change	Ratio of treatment effects:within person variances (on logscale)
NTX	nmol/mmol Cr							
	8	46.7%↓	65.6%↓ to 17.4%↓	41.8%↓	62.5%↓ to 9.9%↓	25.1%	42.5%	0.31
	58	55.3%↓	71.2%↓ to 30.7%↓	51.2%↓	68.5%↓ to 24.4%↓			
	108	62.5%↓	75.8%↓ to 41.9%↓	59.1%↓	73.6%↓ to 36.6%↓			
	158	68.6%↓	79.7%↓ to 51.3%↓	65.7%↓	77.9%↓ to 46.8%↓			
BAP	ng/ml							
	2	2.7%↓	33.2%↓ to 41.9%↑	15.1%↓	41.8%↓ to 23.7%↑	21.2%	25.8%	0.58
	12	13.8%↓	40.9%↓ to 25.6%↑	24.9%↓	48.5%↓ to 9.5%↑			
	22	23.7%↓	47.7%↓ to 11.2%↑	33.5%↓	54.4%↓ to 3.1%↓			

* Mixed model has interactions between treatment and baseline level of marker and between treatment and trial arm.

†Example baseline levels chosen to represent range of data

APPENDIX

Explanation of Model fitting strategy

We fitted a series of mixed models to the FIT bone turn over data where level one was within patient (time) and level two was patient level (all other predictors). A brief summary of model fitting is provided here with an outline of the alternative models in order of increasing complexity. For all models the intercept refers to one year after starting treatment which is the first measurement on treatment. The alternative models were compared using likelihood ratio tests to see which model provided the best fit for the data. Appendix Figure provides a schematic overview of the model fitting process which we used for both NTX and BAP. For both types of models we applied a natural log transformation to marker levels (the outcomes for the models) so that residuals were normally distributed. We started with the least complex model (Model 0) and used forward selection to add the parameters that met the criteria for inclusion until the final model (Model 4) was reached.

The simplest model, Model 0 (random intercept, no treatment effect) included all significant predictors except treatment. For NTX these were: baseline NTX level (nmol/ mmol Cr), baseline hip bone mineral density (g/cm^2), age (years), trial arm (vertebral or clinical fracture arm) and time. For BAP these were: baseline BAP level (ng/ml) and trial arm. The predictors were fitted to have the same effect for everyone (i.e. they were fixed effects). This model also included a term for between-person variation in measurement at one year (random intercept) and a term for within-person variation (residual).

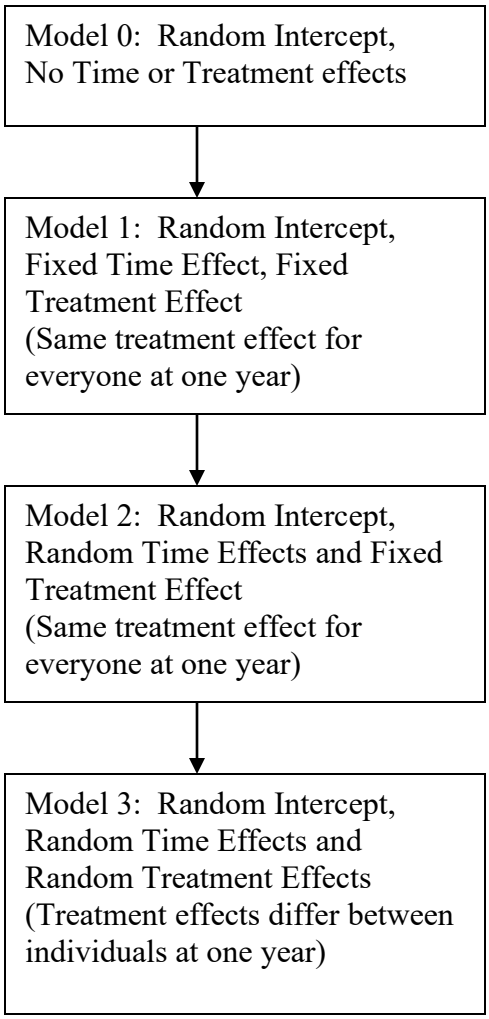
In Model 1 treatment was fitted to have the same effect for everyone (Model 1: random intercept, fixed treatment effect). Model 1 was also used to investigate effect modifiers of treatment. For the NTX models, the following interaction terms were assessed for

significance: treatment×baseline NTX, treatment×baseline hip bone mineral density, treatment×age, treatment×trial arm and treatment×time. For the BAP models, we assessed: treatment×baseline BAP, treatment×trial arm and treatment×time. In this way we identified factors that modify the effect of treatment.

Model 2 was an extension of Model 1 that included a term for between-person variation in change over time to account for the fact that individuals may differ in how their measurements change over time (Model 2: random intercept and random time effects, fixed treatment effect).

Model 3 was an extension of Model 2 that included a term for between-person variation in treatment effect at one year (Model 3: random intercept, random time effects and random treatment effects).

Appendix Figure: Schematic overview of Model fitting (All models have intercepts at 1 year after starting treatment)*



*Adapted from Ref (3). Models were built sequentially from the simplest model (Model 0) to the most complex (Model 3). Likelihood ratio tests were used to decide on the final model (Model 3).

Estimation of number of measurements needed at baseline

We estimated the number of measurements *n* needed so that, for a fixed percentage *y*, the observed average was within *y*% of the true baseline with probability 0.95. In other words, if the true baseline measurement is T_B and the observed mean of n measurements is O_n, then for a fixed percentage *y* we want n such that

$$P\left(\left(1-\frac{y}{100}\right)T_B < O_n < \left(1+\frac{y}{100}\right)T_B\right) \geq 0.95$$

We assumed that that T_B is normally distributed and that $O_n = T_B + (\epsilon_1 + \cdots + \epsilon_n)/n$, where ϵ, the measurement error of a single measurement, is assumed to be normally distributed. We also change the roles of *n* and *y* so that for a fixed number of measurements, we determined y so that with 95% probability the average of the *n* baseline measurements would lie within *y*% of the true baseline. We also repeated all calculations replacing 95% probabilities with 80% probabilities.

For 1 measurement, an observed baseline level of marker would be within 55% of the woman's true baseline level with 95% certainty, or within 34% of the true baseline level with 80% certainty.

Appendix Table 1: Level of uncertainty of observed baseline level of NTX when mean of increasing number of measurements used

Number of measurements used	Percentage of true baseline level that observed level is within, with 95% certainty	Percentage of true baseline level that observed level is within, with 80% certainty
1	55	34
2	38	24
3	31	20
4	27	17
5	24	16
10	17	11
20	12	7

Estimation of number of Measurements needed for estimating apparent change

We use the following method to estimate the number of measurement occasions needed to be certain that if the apparent change in bone turn over marker (A) is < *m*, the true change due to treatment (T) is also < *m*. (We assume here that A and T are the logarithm of the apparent and true change respectively.) The number of measurement occasions needed (*n*) is calculated separately for the time periods before treatment and after treatment has been started. T is assumed to be normally distributed with mean μ and variance ζ^2. The within person variances of the mean of *n* measurement occasions taken before treatment ε_1 and after treatment ε_2 are also assumed to be normally distributed, both with mean 0 and variance σ^2 (mean of n measurements). The apparent change may be written as

$A = T + \varepsilon_1 + \varepsilon_2$, which has a normal distribution with mean μ and variance $\zeta^2 + 2\sigma^2$. The covariance between T and A is ζ^2. To be 95% certain that an apparent change < *m* reflects a true change < *m* mmHg, we seek to find a value of σ^2 (mean of n -of σ^2 (mean of n measurements) such that $P(T < m, A < m) / P(A < m) = 0.95$.

Once we have the value σ^2 (mean of n measurements), the number of measurement occasions (*n)* needed before and after treatment may be calculated from the following equation:

$n = \sigma^2$(1 measurement) /σ^2(mean of n measurements).