PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Study to Weigh the Effect of Exercise Training on BONE quality and strength (SWEET BONE) in type 2 diabetes: Study protocol for a randomized clinical trial
AUTHORS	Balducci, Stefano; Conti, Francesco; Sacchetti, Massimo; Russo, Cosimo; Argento, Giuseppe; Haxhi, Jonida; Orlando, Giorgio; Rapisarda, Gianvito; D'Errico, Valeria; Cardelli, Patrizia; Pugliese, Luca; Laghi, Andrea; Vitale, Martina; Bollanti, Lucilla; Zanuso, S; Nicolucci, Antonio; Pugliese, Giuseppe

VERSION 1 – REVIEW

REVIEWER	Terry Aspray and Antoneta Granic NIHR Newcastle Biomedical research centre, United Kingdom
REVIEW RETURNED	09-Nov-2018

GENERAL COMMENTS	Is the research question or study objective clearly defined?	The study aims to evaluate whether an exercise training programme is effective in improving bone health in older adults with type 2 diabetes. The primary end point is defined (change in the trabecular bone score-TBS) which is a proxy for bone quality. I have concerns about the validity of TBS in the evaluation of bone quality and, in particular the evidence base that *TBS will change under the influence of exercise *Change in TBS for individuals has any established meaning *Degenerative disease (very prevalent in this older population) affecting the spine will not influence the outcomes. Spinal DXA in this age group are often difficult to interpret due to artefact.

	*Power to detect a difference (see statistics section.
2. Is the abstract accurate, balanced and complete?	The abstract seems to present the intended intervention.
3. Is the study design appropriate to answer the research question?	I am unsure about the design for the reasons outlined: *Evidence that change in TBS is detectable and/or meaningful *Size of effect
4. Are the methods described sufficiently to allow the study to be repeated?	Although the intensity and other relevant information about the exercise program (and its components) are described in the Figure 2, the same should be included in the Methods section (Intervention). The authors claim that the feasibility and effectiveness of the program have been tested in a pilot study. However, the intention for this study is quite intensive with a duration of 75 minutes twice a week for over 24 months. The authors should provide more detail about the results of this pilot (number of participants, pilot duration, any adverse events, attrition (and reasons for drop-out), participants' opinions, attitudes, and feedback about the training program (if collected).
5. Are research ethics (e.g. participant consent, ethics approval) addressed appropriately?	Yes
6. Are the outcomes clearly defined?	While the primary and secondary outcomes are clearly defined, participant burden (including the exercise program (Intervention + Standard care) plus PA dairy) is high keeping in mind that the trial will enrol older adults who are inactive and sit more than 8 hours a day.

	7. If statistics are used are they appropriate and described fully?	The power calculation is based on a pilot study comparing the difference between participants with and without diabetes (I assume type 2?). However, that is not the experiment to be undertaken in this study, which intends to compare changes in TBS between adults who are subjected to exercise and those who are not.
	8. Are the references up-to-date and appropriate?	Yes
	9. Do the results address the research question or objective?	For "results" here, I am considering intended outcome measures. There are many "SECONDARY OUTCOMES" While multiple outcome data are to be collected, they may be used in adjusting for outcome data (e.g. using ANCOVA). However, if these are likely to be compared in separate analyses, there should be some discussion of the statistical plan and how statistical significance with multiple comparisons will be addressed. Falls data should be by prospectively collected falls diary, if any meaningful data are to be obtained.
	10. Are they presented clearly?	N/A
	11. Are the discussion and conclusions justified by the results	There is a reasonable introduction, which addresses important aspects of exercise intervention in this group. I think some discussion of potential exercise interventions and intervals and duration of exercise is needed.
	12. Are the study limitations discussed adequately?	Regarding participant burden and secondary outcomes: the participant burden (including the exercise program (Intervention + Standard care) plus PA dairy) is relatively high keeping in mind that the trial will enrol older adults who are inactive and sit more than 8 hours a day. It is to be expected that after 75 minutes of exercise program some muscle soreness and exertion will occur. Did the authors consider including any measures

	of muscle soreness and perceived level of exertion (how physically demanding was the exercise program) and how often such measures would be administered?
13. Is the supplementary reporting complete (e.g. trial registration; funding details; CONSORT, STROBE or PRISMA checklist)?	Yes
14. To the best of your knowledge is the paper free from concerns over publication ethics (e.g. plagiarism, redundant publication, undeclared conflicts of interest)?	Yes
15. Is the standard of written English acceptable for publication?	Yes

REVIEWER	Stavroula A. Paschou Medical School, National and Kapodistrian University of Athens, Athens, Greece
REVIEW RETURNED	16-Dec-2018

GENERAL COMMENTS	Very well performed and written

REVIEWER	Judith Godin Dalhousie University Canada
REVIEW RETURNED	15-Mar-2019

GENERAL COMMENTS	I was asked to do a statistical review and the subject matter is outside of my expertise. Overall, I found the paper well structured and written with an appropriate level of detail. I would recommend a reduction in the number of acronyms to improve ease of reading. I have the following considerations largely in relation to the proposed statistical analyses: • I commend the authors for having a strategy in place to minimize the amount of missing data.

	• When using notation such as 1.225 ±0.085, it would be helpful to specify what the value after the ± represents (e.g., SD, SE) at first use. • It appears the power calculations are based on a one-tailed test and the smaller of the two SDs obtained from the pilot study. Consider providing more information and rationale regarding the power calculation. Also, you may want to consider whether you will have enough power for your sensitivity analysis and your subgroup analysis. • Consider a two-way mixed ANOVA (Time X Intervention). If the interaction between time and intervention is statistically significant, you could examine how changes over time were different between the intervention and control groups. This may require an additional power analysis. • It is not clear that the intervention will be included as an IV in the multiple regression, but I'm assuming that this is the case. • I'm not familiar with the method suggested for dealing with missing data. A brief explanation of the technique and rationale would be helpful.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
The study aims to evaluate whether an exercise training programme is effective in improving bone health in older adults with type 2 diabetes. The primary end point is defined (change in the trabecular bone score-TBS) which is a proxy for bone quality. I have concerns about the validity of TBS in the evaluation of bone quality.
The primary objective of our study was to assess the effect of an exercise intervention program on bone quality in patients with type 2 diabetes (T2D), based on the assumption that a reduced bone quality is the main cause for the increased fracture risk in these individuals, who have preserved bone mass. Among the various measures of bone quality, we chose TBS as the primary end-point, because "it was consistently found to be reduced in T2D patients with versus those without fracture and in T2D versus non-diabetic individuals, and to predict fractures independently of BMD". If not TBS, which other measure of bone quality should we use? Is there any other measure which the Reviewers consider valid in the evaluation of bone quality?
In particular the evidence base that:
a. TBS will change under the influence of exercise;
b. Change in TBS for individuals has any established meaning;
c. Degenerative disease (very prevalent in this older population) affecting the spine will not influence the outcomes. Spinal DXA in this age group are often difficult to interpret due to artefact.
d. Power to detect a difference (see statistics section).
As stated in the manuscript, "the SWEET-BONE is the first study investigating whether a specifically designed exercise training program is effective in improving bone quality and strength in patients with T2D, thus potentially reducing the increased fracture risk characterizing these individuals despite preserved bone mass" (please, see page 3, line 2, and page 18, line 11). Therefore, there are no data on whether TBS or any other measure of bone quality change with exercise, except for a few findings from animal studies (please, see Fonseca H et al. Sports Med. 2014;44:37-53), or on the meaning of this change. Our study was specifically designed to provide these data, by assessing whether or not:
1. TBS (primary) and other parameters of bone quality (secondary) improve with exercise training;

2. an improvement in these parameters (if observed) translates into a reduced fracture risk over an extended 7-year follow-up.

We do not understand why the lack of data on these issues should be a matter of concern. Several studies have previously investigated the effect of exercise training on bone mass, even if it was unclear whether "BMD changes under the influence of exercise" and whether "change in BMD for individuals has any established meaning". As recently reviewed (please, see Benedetti MG et al. Biomed Res Int. 2018;2018:4840531), these studies showed that exercise is effective in increasing BMD, especially weight-bearing aerobic exercises, strength and resistance exercises, and multicomponent training program containing these exercises. In addition, several studies showed that exercise prevents BMD reduction during dietary weight loss in obese older adults (please, see Villareal et al. N Engl J Med. 2017;376:1943-1955), including those with diabetes (please, see Daly RM et al. Osteoporos Int. 2005;16:1703-1712), though other surveys failed to detect a significant effect (please, see Beavers DP et al. Osteoarthritis Cartilage. 2014;22:726-733). Nevertheless, it is still unknown whether exercise-induced increase in BMD results in reduced risk of fracture. Therefore, we do believe that it would be worth to evaluate also the effect of exercise on bone quality and the associated fracture risk and that lack of data is one more reason to perform these studies. In addition, we do not understand whether the Reviewers' critique is restricted to TBS or extended to all the other measures of bone quality, which would suffer from the same limitation, as no study has assessed the effect of exercise on these parameters and the consequent impact on bone fragility.

We do agree with the Reviewers that degenerative disease is very prevalent in this older population; nevertheless, studies with spinal DXA are conducted mainly in this age group that carries an increased fracture risk. In our study, vertebral morphometry will be performed in combination with DXA, thus allowing to identify individuals with degenerative disease which might influence outcomes. In addition, patients with spinal deformity index higher than 5 (and higher than 2 in a single vertebra) will be excluded from the study. Finally, and more importantly, there is evidence that, at variance with BMD (please, see Liu G et al. Osteoporos Int. 1997;7:564–569), TBS is not affected by lumbar spine osteoarthritis, because it is determined by local gray-level variations, and not by their absolute levels (please, see Kolta S et al. Osteoporos Int. 2014;25:1759–1764). For this reason, TBS is now preferred over BMD for evaluating bone status and fracture risk in individuals with spondyloarthritis (please, see Martineau P & Leslie WD. Bone. 2017;104:66–72; Boussoualim K et al. Joint Bone Spine. 2018;85:727–731; and Kang KY et al. Rheumatology. 2018;57:1033-1040). In any case, as TBS is calculated from the lumbar spine region used for BMD calculation, deleting vertebrae which would not be used for BMD estimation automatically removes them from the TBS calculation.

Regarding the power to detect a difference, please, see below.

I am unsure about the design for the reasons outlined:

- Evidence that change in TBS is detectable and/or meaningful;
- Size of effect.

Please, see responses to the comments above.

Although the intensity and other relevant information about the exercise program (and its components) are described in the Figure 2, the same should be included in the Methods section (Intervention). The authors claim that the feasibility and effectiveness of the program have been tested in a pilot study. However, the intention for this study is quite intensive with a duration of 75 minutes twice a week for over 24 months. The authors should provide more detail about the results of this pilot (number of participants, pilot duration, any adverse events, attrition and reasons for drop-out, participants' opinions, attitudes, and feedback about the training program, if collected).

All the relevant information about the exercise program have been provided in Figure 2 and not in the Methods section due to space constraints. Following to the Reviewers' suggestion, we have now spared space somewhere else in order to expand the description of the training program in the main text (please, see page 11, line 16). As presented in the Methods section, Figure 2 and the figure legend, the intervention is a "progressive" training program, starting at low intensity and gradually progressing to moderate intensity, with adjustment of intensity according to improvements in physical fitness (i.e., cardiorespiratory and muscular fitness) and gradual increase in the difficulty level and weight of vests. In more detail:

- the intensity of aerobic exercise will be increased from 50% VO2max to 70% VO2max at the end of year 1 and will be adjusted according to improvements in predicted VO2max, as recorded every 6 months;
- the intensity of resistance exercise will be increased from 60% 1-RM to 80% 1-RM at the end of month 6 and will be adjusted according to improvements in in 1-RM, as recorded every 6 months, and the velocity of execution during the concentric phase of the movement will be progressively increased;
- the impact of weight bearing exercises will be increased from light (stationary movements, i.e., stomping, jumps), to moderate (forward and backward movements, i.e., box step-ups, pogo jumps), and high (multilateral/ multidirectional movements, i.e., side-to-side shuffle, lateral box jumps) and the height of jumps and amplitude of movements will be progressively increased;
- the difficulty level of balance training will be gradually increased by performing the exercises with closed eyes, reducing the support area, changing visual fixation (e.g., head rotations), varying the centre of mass (e.g., limb raising), or adding a manual or cognitive task;
- the weighted vests will be worn only after the first month, during each session (while performing aerobic training and "weight bearing" exercises) and also outside the sessions (at least 1 hour plus 3 10-repetition series of step-up and sit-to-stand in three non-training days every week) and weight of vests will be increased by 2% of body weight every 6 months (i.e., from 2% to 8%).

Progression of exercise intensity, difficulty level of balance training, and weight of vests ensures safety and prevents attrition (please, see page 11, line 23), thus making the intervention feasible. Indeed, feasibility of twice-a-week (or even more frequent) training sessions in sedentary and physically inactive elderly individuals has already been demonstrated by several studies from our group and other investigators. In the IDES, about 1/3 of the 303 participants randomized to twice-a-week, 75-min sessions of progressive aerobic and resistance exercise training for one year were older than 65 years, i.e., in the same age range of the SWEET-BONE participants or even older (some of them were 76 to 80 years old); nevertheless, they did not experience major adverse events during the sessions and drop-out was very low and unrelated to the exercise program (please, see Balducci S et al. Arch Intern Med. 2010;170:1794–1803). As stated in the manuscript, the pilot study was performed to set-up the specific training program for improving bone health (please, see page 11, line 13) and also to train physicians, exercise specialists, and outcome assessors participating in the trial (please, see page 9, line 2). In addition, though this was not the scope of the pilot study, it allowed us to preliminarily evaluate both the efficacy and safety of the intervention. It was conducted on 20 T2D patients meeting the inclusion/exclusion criteria of the SWEET-BONE and lasted one year. There were no major adverse events and no drop-outs.

While the primary and secondary outcomes are clearly defined, participant burden (including the exercise program (Intervention + Standard care) plus PA dairy) is high keeping in mind that the trial will enrol older adults who are inactive and sit more than 8 hours a day.

Please, see responses to the comments above.

The power calculation is based on a pilot study comparing the difference between participants with and without diabetes (I assume type 2?). However, that is not the experiment to be undertaken in this study, which intends to compare changes in TBS between adults who are subjected to exercise and those who are not.

As stated above (and in the manuscript), this is the first study assessing the effect of exercise training on TBS and other measures of bone quality in T2D individuals. In addition, in the pilot study, we compared 40 T2D and 20 non-diabetic individuals at baseline, then 20 of the T2D patients participated in a one-year exercise program and no one of them served as non-training control. Thus, there are no data available from studies comparing the effect of exercise versus no exercise on TBS on which to base power calculation. For this reason, we calculated the sample size required to observe an improvement in the exercise group which allows to bridge the gap of 0.030 with non-diabetic controls detected in the pilot study at baseline (1.225 vs 1.255) with a statistical power of

90% (α=0.05) by unpaired t test. We have now recalculated the sample size (please, see page 16, line 7) based on the baseline TBS levels detected in the T2D individuals from our pilot study (1.225±0.085) and an effect size of 0.50. To detect a between group difference of 0.045 in TBS (i.e. an effect size of 0.50) with a statistical power of 90% (α=0.05) by two-sided two-sample equal-variance t-test, 86 patients per arm are needed (172 total). A sample of 200 patients allows to tolerate a 14% dropout rate.

For “results” here, I am considering intended outcome measures. There are many “SECONDARY OUTCOMES”. While multiple outcome data are to be collected, they may be used in adjusting for outcome data (e.g. using ANCOVA). However, if these are likely to be compared in separate analyses, there should be some discussion of the statistical plan and how statistical significance with multiple comparisons will be addressed. Falls data should be by prospectively collected falls diary, if any meaningful data are to be obtained.

We agree with the Reviewer that there are many secondary endpoints, which however include all the main parameters that can be affected by the exercise training program and may in turn affect fracture risk, in addition to TBS, i.e., other potential measures of bone quality, as assessed by quantitative ultrasound and peripheral quantitative computer tomography; bone mass; markers of bone turnover; muscle strength, mass, and power; balance and gait; and, in the extended follow-up, falls and asymptomatic and symptomatic fractures. We also agree that this makes concern about type 1 error important, though it is still debated whether or not adjustment for multiple comparison is appropriate in these circumstances (please, see Bender R & Lange S. J Clin Epidemiol. 2001; 54:343–349). We prefer to interpret findings for analyses of secondary endpoints as exploratory (please, see page 17, line 4), as done in a recent publication from our group (please, see Balducci S. et al. JAMA. 2019;321:880-890). Regarding the fall questionnaire, this is a validated tool for assessing falls (please, see ref #56), widely utilized in studies on bone fractures, and represents the only feasible method to collect information of this issue.

There is a reasonable introduction, which addresses important aspects of exercise intervention in this group. I think some discussion of potential exercise interventions and intervals and duration of exercise is needed.

Please, see responses to the comments above.

Regarding participant burden and secondary outcomes: the participant burden (including the exercise program (Intervention + Standard care) plus PA dairy) is relatively high keeping in mind that the trial will enrol older adults who are inactive and sit more than 8 hours a day. It is to be expected that after 75 minutes of exercise program some muscle soreness and exertion will occur. Did the authors consider including any measures of muscle soreness and perceived level of exertion (how physically demanding was the exercise program) and how often such measures would be administered?

We have extensively addressed the issue of the participant burden in the above responses to the Reviewers’ comments. Measures of muscle soreness and perceived level of exertion are semiquantitative parameters with limited validity, which would only increase the number of secondary endpoints. However, as now stated in the Patient and Public Involvement section (please, see page 17, line 7), prior to starting each session, patients are invited to report to the exercise specialist any symptom that may hamper or limit participation, including muscle soreness and perceived level of exertion, as in previous studies from our group (please, see Balducci S et al. Acta Diabetol. 2014;51:647-654). This allows the exercise specialist supervising the session to identify the appropriate exercise types and modalities in order to avoid symptom exacerbation and minimize the risk of injury or adverse events (e.g., temporary reduction of the exercise intensity, in case of muscle soreness, or temporary exclusion of painful segments, in case of pain).

Reviewer: 2
Very well performed and written.
We thank the Reviewer for her positive judgement on our work.

Reviewer: 3
I was asked to do a statistical review and the subject matter is outside of my expertise. Overall, I found the paper well-structured and written with an appropriate level of detail. I would recommend a reduction in the number of acronyms to improve ease of reading.
We thank the Reviewer for her favourable comments and useful suggestions.
I commend the authors for having a strategy in place to minimize the amount of missing data.
We thank the Reviewer for outlining the importance of minimizing the amount of missing data,
When using notation such as 1.225±0.085, it would be helpful to specify what the value after the ± represents (e.g., SD, SE) at first use.
It is SD, as now specified in the revised version (please, see page 16, line 7).
It appears the power calculations are based on a one-tailed test and the smaller of the two SDs obtained from the pilot study. Consider providing more information and rationale regarding the power calculation. Also, you may want to consider whether you will have enough power for your sensitivity analysis and your subgroup analysis.
We based our original power calculation on the higher (0.085), not the smaller (0.067) SD, because the former was from T2D patients whereas the latter was from non-diabetic individuals. Regarding the rationale of power calculation, there are no data available from studies comparing the effect of exercise versus no exercise on TBS on which to base power calculation, as this is the first study addressing this issue. Moreover, in the pilot study, we compared 40 T2D and 20 non-diabetic individuals at baseline, then 20 of the T2D patients participated in a one-year exercise program and no one of them served as non-training control. For this reason, we calculated the sample size required to observe an improvement in the exercise group which allows to bridge the gap of 0.030 with non-diabetic controls detected in the pilot study at baseline (1.225 vs 1.255) with a statistical power of 90% (α=0.05) by unpaired t test. Following to a comment of other Reviewers, we have now recalculated the sample size (please, see page 16, line 7) based on the baseline TBS levels detected in the T2D individuals from our pilot study (1.225±0.085) and an effect size of 0.50. To detect a between group difference of 0.045 in TBS (i.e. an effect size of 0.50) with a statistical power of 90% (α=0.05) by two-sided two-sample equal-variance t-test, 86 patients per arm are needed (172 total). A sample of 200 patients allows to tolerate a 14% dropout rate. Regarding the power for the sensitivity and subgroup analyses, power is usually calculated only for the primary analysis in RCTs.
Consider a two-way mixed ANOVA (Time X Intervention). If the interaction between time and intervention is statistically significant, you could examine how changes over time were different between the intervention and control groups. This may require an additional power analysis.
The duration of the follow-up is the same for all participants (i.e., 2 years) and time is already accounted for by mixed models for repeated measures.
It is not clear that the intervention will be included as an IV in the multiple regression, but I'm assuming that this is the case.
Yes, correct.
I'm not familiar with the method suggested for dealing with missing data. A brief explanation of the technique and rationale would be helpful.
As, now reported in the Methods section (please, see page 16, line 23), models for repeated measures with an autoregressive correlation type matrix make an assumption of missing at random and account for both missingness at random and potential correlation within participants, as they allow evaluating all individuals, including those with incomplete data (please, see Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer, 2000; and Singer JD, Willett JB. Applied longitudinal data analysis: modeling change and event occurrence. New York: Oxford University Press, 2003).

VERSION 2 – REVIEW

REVIEWER	Dr Terry Aspray Newcastle University
REVIEW RETURNED	05-May-2019

GENERAL COMMENTS	I believe the comments of reviewer 1 have yet to be addressed, despite a robust argument being presented by the authors. The main scientific concern is the use of a change in DXA TBS (DeltaTBS) as a primary outcome. While the authors may want to evaluate bone "quality" - whatever that means as it is not defined in the manuscript, there remains no useful clinical evidence for this or any other measure. 1. We do not know whether there is a DeltaTBS for normal or type 2 diabetic populations- If there is one what is the expected size (without the impact of exercise)? At least none is presented here 2. We do not know the meaning of a change in TBS - e.g. even from trials of antiresorptive agents, Leslie et al have shown DeltaTBS has no clinical relevance to fracture risk and, they state "unlike antiresorptive treatment-related changes in BMD, change in lumbar spine TBS is not a useful indicator of fracture risk irrespective of osteoporosis treatment" (JBMR, 2017 DOI 10.1002/jbmr.3054). At the very least, these issues should be discussed in the ms. The statistical reviewer admitted the subject matter is outside their expertise. I remain uncomfortable with the use of DeltaTBS as the primary outcome with no pilot data presented for this and so no way to estimate the mean expected change in TBS or its variance (SD). I hope that the statistician is aware of this in their comments. Overall, i still find this paper unacceptable for publication. I fear that the authors must feel me antagonistic and the tone of the reply to my first comments suggests a lack of understanding of the issues I am raising. However, my comments are those of a clinical trialist, an experienced reviewer and associate editor for a biomedical journal. My wish is that the authors get the design of their study and the outcome measures clearly justified and presented before embarking on a study using an (as yet unproven) biomarker.

REVIEWER	Judith Godin Dalhousie University and Nova Scotia Health Authority Canada
REVIEW RETURNED	06-May-2019

GENERAL COMMENTS	The authors have addressed my concerns.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

I believe the comments of reviewer 1 have yet to be addressed, despite a robust argument being presented by the authors.

The main scientific concern is the use of a change in DXA TBS (DeltaTBS) as a primary outcome. While the authors may want to evaluate bone "quality" - whatever that means as it is not defined in the manuscript, there remains no useful clinical evidence for this or any other measure.

1. We do not know whether there is a DeltaTBS for normal or type 2 diabetic populations- If there is one what is the expected size (without the impact of exercise)? At least none is presented here.

2. We do not know the meaning of a change in TBS - e.g. even from trials of antiresorptive agents, Leslie et al have shown DeltaTBS has no clinical relevance to fracture risk and, they state "unlike antiresorptive treatment-related changes in BMD, change in lumbar spine TBS is not a useful indicator of fracture risk irrespective of osteoporosis treatment" (JBMR, 2017 DOI 10.1002/jbmr.3054).

At the very least, these issues should be discussed in the ms.

The statistical reviewer admitted the subject matter is outside their expertise. I remain uncomfortable with the use of DeltaTBS as the primary outcome with no pilot data presented for this and so no way to estimate the mean expected change in TBS or its variance (SD). I hope that the statistician is aware of this in their comments.

Overall, I still find this paper unacceptable for publication. I fear that the authors must feel me antagonistic and the tone of the reply to my first comments suggests a lack of understanding of the issues I am raising. However, my comments are those of a clinical trialist, an experienced reviewer and associate editor for a biomedical journal. My wish is that the authors get the design of their study and the outcome measures clearly justified and presented before embarking on a study using an (as yet unproven) biomarker.

It appears that we have successfully addressed all the criticisms raised by Reviewer #1 (e.g., influence of degenerative disease on trabecular bone score [TBS], detailed information about the exercise intervention, participant burden from the training program, multiple secondary outcomes, and fall questionnaire), except change in TBS (DeltaTBS) as primary outcome of our study, which is still a matter of concern.

Before further discussing this issue in detail, we want to spend a few words on the final comments of Reviewer #1. Rather than claiming a lack of understanding of anyone, we prefer to say that we and Reviewer #1 disagree with each other. This does not mean that we question the Reviewer's expertise as a clinical trialist, reviewer and journal editor. Thus, we do understand the Reviewer's point of view but, based on current evidence, we have a different opinion, as discussed below. However, the issue here is not if we should embark in this trial, but if we have clearly justified and presented the study design and outcomes, thus making our manuscript suitable for publication. We believe that we have already provided a clear justification for focusing our study on bone quality and, as a consequence, for choosing DeltaTBS as primary outcome but, as the Reviewer is not satisfied, we will try again to explain the rationale for the study design and outcomes by further discussing the issues regarding bone quality and TBS.

Bone quality

The Reviewer thinks that it is not worth to assess the effect of an intervention on such an unclear entity (“whatever that means”) and that “it is not defined in the manuscript”.

As clearly stated in the manuscript and in our previous response to Reviewer #1, the rationale for focusing on “bone quality” is that, in patients with type 2 diabetes (T2D), the increased fracture risk has been attributed to a reduced bone quality, since these individuals present with preserved or even increased bone mass. That T2D is associated with altered material and structural properties resulting in poor bone quality and strength has long been recognized and is now widely accepted (see Schwartz AV. Calcif Tissue Int. 2003;73:515–519; Leslie WD et al. J Bone Miner Res. 2012;27:2231–2237; Farr JN & Khosla S. Bone. 2016;82:28–34). This is unlike individuals with osteoporosis, who show reduction of both mass and quality (though quality is reduced to a lesser extent than in T2D patients). In addition, it is already known that exercise, especially resistance training, is effective in improving bone mineral density (BMD) in post-menopausal women with reduced bone mass (see ref. #38), though it is unclear whether it may affect BMD also in individuals with normal-to-increased bone mass. However, DeltaBMD is a secondary endpoint of our study, thus allowing comparison of the effects of exercise on quality (TBS and other measures) and mass (BMD).

A “definition of bone quality” is provided in all the three versions of the manuscript, i.e., the first two already evaluated by the Reviewer and the current one. In the Introduction section (see page 5, line 10), it is clearly stated that “Bone quality is determined by (a) bone architecture, including geometry (macro-architecture) and micro-architecture; and (b) material properties, including mineralization and collagen cross-links, which in turn are influenced by bone turnover as well as by accumulation of microdamage and microstructural discontinuities such as microporosity and lamellar boundaries”. In our opinion, this is a definition of bone quality, which is widely accepted by the scientific community. For instance, according to the 2015 official position of the International Society for Clinical Densitometry (ISCD) “A number of skeletal features other than BMD, such as bone geometry, microarchitecture, mineralization, bone remodeling, and microdamage contribute to bone strength and overall fracture risk. These features and characteristics of the skeleton that influence a bone’s ability to resist fracture are known as bone quality” (see Silva BC et al. J Clin Densitom. 2015;18:309-330). Based on this definition, hip geometry, cortical porosity, and TBS can be considered as measures of bone quality (see below for TBS).

TBS

The Reviewer thinks that “this or any other measure” of bone quality are not suitable as study outcomes as there is “no useful clinical evidence”. In particular, he believes that:

1. “it is unknown whether there is a DeltaTBS for normal or type 2 diabetic populations and, if there is one, what is the expected size (without the impact of exercise), “at least none is presented here” and, hence, “there is no way to estimate the mean expected change in TBS or its variance” thus precluding sample size calculation.

2. “we do not know the meaning of a change in TBS” and even from trials of antiresorptive agents Leslie et al showed that “DeltaTBS has no clinical relevance to fracture risk”.

As already discussed in our previous response to Reviewer #1, since we “want to evaluate bone quality” for the reasons outlined above, we believe that TBS is the best proxy to use because it is considered a reliable index of trabecular microarchitecture and, hence, of bone quality (see Silva BC et al. J Clin Densitom. 2015;18:309-330). In fact, though TBS is not a direct physical measurement of bone microarchitecture (see Silva BC et al. J Bone Miner Res. 2014;29:518-530), it correlates with trabecular microarchitecture parameters, including connectivity density, trabecular number, and trabecular separation, as shown by a number of ex vivo studies (see Hans D et al. J Clin Densitom. 2011;14:302-312; Winzenrieth R et al. J Clin Densitom. 2013;16:287-296; Roux JP et al. Osteoporos Int. 2013;24:2455-2460). In addition, TBS was consistently found to be reduced in T2D patients with fracture versus those without (see ref. #14) and in T2D versus non-diabetic individuals (see refs #15-20 and a recent meta-analysis by Ho-Pham LT & Nguyen TV. Osteoporos Int. 2019 Jun 18), and to predict fracture risk independently of BMD (see ref. #15).

As also stated in our previous response to Reviewer #1 and in the manuscript (see page 16, line11), "the SWEET-BONE is the first study investigating whether a specifically designed exercise training program is effective in improving bone quality and strength in patients with T2D, thus potentially reducing the increased fracture risk characterizing these individuals despite preserved bone mass". Therefore, "there are no data on whether TBS or any other measure of bone quality change with exercise, except for a few findings from animal studies (see Fonseca H et al. Sports Med. 2014;44:37-53), or on the meaning of this change" and "our study was specifically designed to provide these data, by assessing whether or not (1) TBS (primary) and other parameters of bone quality (secondary) improve with exercise training; and (2) an improvement in these parameters (if observed) translates into a reduced fracture risk over an extended 7-year follow-up". Thus, we have already agreed with Reviewer #1 that there are no data on DeltaTBS in T2D individuals with exercise and, hence, that training may or may not produce a significant change in TBS versus standard care, as a result of an increment, a stabilization, or a lower decline in the intervention group compared to the expected time-dependent reduction in the control group. In addition, due the lack of data on the mean expected change in TBS or its variance, sample size was estimated from the baseline data of our pilot study, i.e., from the gap in TBS between T2D and non-diabetic controls (1.225 vs 1.255) and, then, from the TBS levels in T2D individuals (1.225±0.085) and an effect size of 0.50. Reviewer #3 (the statistical reviewer) was satisfied with this latter estimate. However, if it is true that there are no data on TBS change in T2D patients and in response to exercise, it is not true that "there are no data on TBS change with time in non-diabetic individuals" and that "change in lumbar spine TBS is not a useful indicator of fracture risk irrespective of osteoporosis treatment".

Regarding "DeltaTBS in non-diabetic individuals", TBS was found to decrease by 0.4-0.5% per year in large cross-sectional studies in French (see Dufour R et al. Osteoporos Int. 2013;24:2837-2846) and Japanese (see Iki M et al. Osteoporos Int. 2015;26:1841-1848) women. Moreover, TBS was found to decrease by approximately 0.3% per year in a longitudinal study in older women from the Manitoba Bone Density Program (Krieg MA et al. Osteoporos Int. 2013;24:1073-1078). Finally, even the same paper cited by Reviewer #1 reported that, over a mean follow-up of 4 years, mean reduction in TBS was 1.2% vs a 1.7% decrease in spine BMD in untreated women from the Manitoba Bone Density Program (see Leslie WD et al. J Bone Miner Res. 2017;32:618-623). Though the extent of the age-related decrease in TBS is unknown in the diabetic population, even without the impact of exercise, it is likely to be greater than in the general population, given the large decrease observed in T2D vs non-diabetic individuals, suggesting that the diabetic condition accelerates the age-dependent TBS reduction.

Regarding "DeltaTBS with osteoporosis treatment", as correctly reported by Reviewer #1, Leslie et al did "conclude that, unlike antiresorptive treatment–related changes in BMD, change in lumbar spine TBS is not a useful indicator of fracture risk irrespective of osteoporosis treatment", based on the finding that TBS increased by 0.8% compared with a 6.3% increment in spine BMD (see Leslie WD et al. J Bone Miner Res. 2017;32:618-623). Surprisingly, Reviewer #1 omitted another important comment by Leslie et al, i.e., that "our data may not apply to osteoanabolic therapies that generate more robust changes in both lumbar spine BMD and TBS", a statement supported by 3 references (see Senn C et al. Osteoporos Int. 2014;25:1945–1951; Di Gregorio S et al. Bone. 2015;75:138–143; Saag KG et al. Arthritis Rheumatol. 2016;68:2122–2128). In these studies, the increase in TBS with teriparatide (~4% in 2-3 years) was much larger than that with antiresorptive agents, though lower than that in spine BMD. A meaningful increase in TBS was reported also in more recent studies with teriparatide (see Miyaoka D et al. Calcif Tissue Int. 2017;101:396–403; Tsai JN et al. J Clin Densitom. 2017;20:507–512; Ebina K et al. J Bone Miner Metab. 2018;36:478–487) and abaloparatide (Bilezikian JP et al. Osteoporos Int. 2018;29:323–328). This impressive difference in TBS responsivity to drug treatments can be explained by the robust evidence that this measure reflects trabecular microarchitecture (see above). In fact, the lack of usefulness of TBS could be anticipated by the mechanism of action of antiresorptive drugs, as they merely increase bone mineralization, i.e., calcium content and ultimately bone mineral content (BMC)/BMD. These changes, however, do not affect trabecular microarchitecture to any significant extent and, therefore, produce only a small effect on TBS, unlike osteoanabolic agents, which have the potential to improve microarchitecture. Though there are no data on the effect of exercise on TBS change, it is well known that exercise, depending on type, load and other factors that we have considered in designing our intervention protocol, has a potential osteoanaboolic effect on bone, and is therefore likely to influence positively microarchitecture and TBS (see Russo CR. Clin Cases Miner Bone Metab. 2009;6:223-228). This concept is consistent with a recent cross-sectional study from the NHANES 2005–2006 showing that older, but not younger, women and men with higher levels of activity, as measured objectively by an accelerometer, had higher BMD and TBS, and that benefits were noted with as little as 5–20 min of daily physical activity (see Jain RK & Vokes T. Arch Osteoporos. 2019;14:29).

All the above considerations are summarized in the 2019 official position of the ISCD stating that "BMD measures bone quantity and TBS measures bone quality. These tests can be considered complementary in assessing fracture risk and response to therapy in appropriately selected patients. A statistically significant decrease of TBS in a treated or untreated patient may represent a clinically relevant worsening of trabecular structure and increasing fracture risk" (see Krohn K et al. J Clin Densitom. 2019 Jul 9).

Finally, we do agree with the Reviewer that the issue of DeltaTBS was not discussed in detail in the manuscript, but this was due to space constraints. Following to the Reviewers' suggestion, we have now created space elsewhere in order to expand these issue as far as possible. Specifically, while the definition of bone quality was already in the manuscript (see page 5, line 10), we added a paragraph regarding TBS change stating that "Potential pitfalls include the lack of data on TBS change over time in T2D individuals and the impact of exercise on this surrogate measure of bone quality. However, an age-dependent reduction in TBS of up to 0.5%/year has been reported in the general population and such decrease is likely to be accelerated in T2D patients, given the large reduction in TBS detected in T2D versus non-diabetic individuals. In addition, in osteoporotic individuals, TBS was shown to be markedly increased (by ~4% in 2-to-3 years) by osteoanabolic agents such as teriparatide, though less than spine BMD, whereas antiresorptive agents, which merely increase bone mineralization, were virtually ineffective. Therefore, exercise, by virtue of its potential osteoanaboolic effect, is likely to influence positively TBS, consistent with a recent cross-sectional study showing that people with higher levels of objectively measured PA had higher TBS (and BMD)" (see page 16, line 15).

Reviewer: 3

The authors have addressed my concerns.

We thank the Reviewer for her positive comments.