Peer Review File

Article information: https://dx.doi.org/10.21037/atm-23-1699

Reviewer A

Comment 1: Line 70: There are other groups that already developed finite element models which are used for fracture prediction in patients (Goodheart et al, Eggermont et al, Sternheim et al). Why would it be preferred to do the tests of this study in a canine osteosarcoma model instead of a human osteosarcoma model?

Reply 1: The authors agree that previous models have been generated using human osteosarcoma models, which are ultimately more relevant to human healthcare. However, many previous models have not demonstrated large-scale clinical validation. Due to the relatively greater prevalence of osteosarcomas in canines as compared to humans, our intent was to leverage this patient population as a means of expedited clinical validation of our fracture prediction model. We wanted to start with a canine osteosarcoma model as a preliminary study to validate the procedure before moving on to testing with lesions and human tissue. The benefit of using the canine model is that we could eventually test the modeling procedure on living canine patients to determine whether it's clinically feasible; this would allow us to make the appropriate changes to the technique necessary to be useful in humans. Future work will include clinical validation of this model and confirmation of translatability to humans.

Changes in the text: We have added text to the Introduction to explain our rationale for using a canine model, as this was not clear in the original manuscript (Page 5, line 90). "Many previous models have not demonstrated large-scale clinical validation. Due to the relatively greater prevalence of osteosarcoma in canines (15) as compared to humans, a canine model was used in this study to leverage the patient population as a means of expedited clinical validation of the fracture prediction model. The canine osteosarcoma model will serve as a preliminary model to validate the procedure before moving on to testing with lesions and human tissue. The benefit of starting with canines is that the modeling procedure could eventually be tested on living canine patients to determine whether it's clinically feasible and allow for making appropriate changes to the technique necessary to be useful in humans."

Comment 2: Line 97/line 274: Is the semicircular osteotomy a good representation of an osteosarcoma? What differences would you expect between the osteotomy and an action osteosarcoma lesion?

Reply 2: Thank you for the questions. We would expect an osteosarcoma lesion to have a different macroscopic structure and material properties from the osteotomy. The purpose of the osteotomy was to simulate weakening of the bone, but it does not encapsulate the aforementioned structural and material differences. To address this, future work will include validating the modeling technique on bones with osteosarcoma lesions.

Changes in the text: We have modified our text as advised to explain the limitations of using an osteotomy rather than actual lesions (Page 18, line 347). "Although the osteotomy simulates the weakening of the bone that occurs in active lesions, it does not take into account structural and material properties of the tumor. Future work will further investigate the use of FE models on limbs with clinically diagnosed osteosarcoma to better represent the effect of tumors on fracture mechanics."

Comment 3: Line 102: The different loading conditions are referred to as "treatment group", which was confusing. Using the term "loading group/loading condition" could clarify this.

Reply 3: We agree that this terminology was misleading and have changed it to clarify the groups.

Changes in the text: We have modified our text as advised (Page 6, line 115). "… were randomly allocated to one of three loading groups for biomechanical testing."

Comment 4: Line 103: Only a few samples are used for each of the loading conditions. What is the power of this study?

Reply 4: The authors appreciate this comment. For this initial study, we had a limited number of samples available, and wanted to make sure the model could predict responses in all three loading conditions (compressions, bending, and torsion). We ran a statistical power analysis, which resulted in power values below 50%. Because of the low sample size, we regonize that statistical analysis may not be the most appropriate method of analysis. For this reason, we decided to switch to a linear regression to compare experimental and computational results, which also allowed for more direct comparison between similar published studies.

Changes in the text: We modified our text as advised by acknowledging low sample size as a limitation (Page 17, line 331). "The main limitation was that this study had a small sample size; additional samples are needed in future studies to support validation of the models more robustly."

We also added a linear regression analysis in place of statistical and power analysis (Page 14, line 247). "Linear regressions were performed comparing experimental and finite element yield loads to determine the accuracy of the FE predictions. Plots of the regression are shown in Figure 7 for bending and compression (Figure 7A) and for torsion (Figure 7B). The respective R^2 values were 0.9335 and 0.8798."

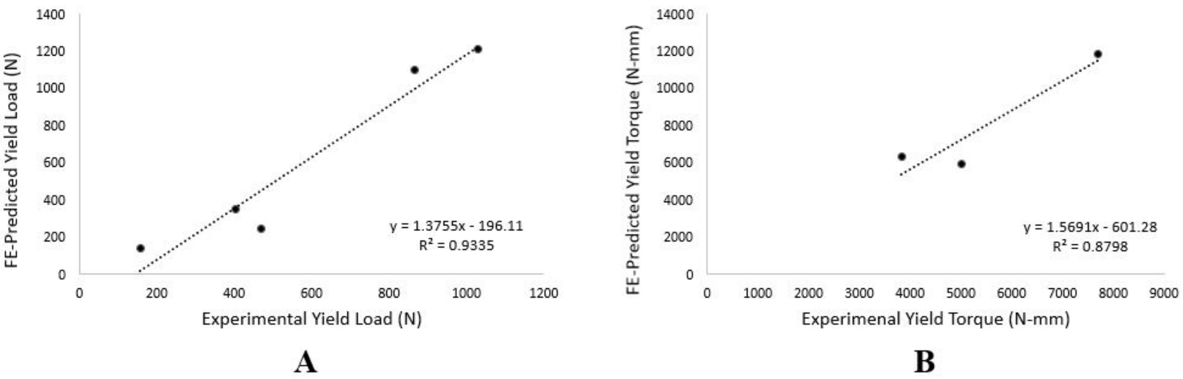


Figure 7: Linear regression comparing experimental and finite element yield for compression and bending (A) and torsion (B). Compression and bending were grouped together to ensure sample size was sufficient to perform linear regression.

Comment 5: Line 156: "…, medullary cavity elastic modulus was manually input as a constant isotropic value of 0.02 MPa". How was the medullary cavity determined/segmented?

Reply 5: The authors agree that the medullary cavity segmentation criteria was unclear in the original submission. The medullary cavity was defined using a Houndsfield (HU) cutoff for bone, and regions with a lower HU were considered part of the medullary cavity. Previous studies have investigated HU ranges for bone (such as the Kim et al. paper "Houndsfield units on lumbar computed tomography for prediciting regional bone mineral density"). We used a value of 226 HU as the lower bound for bone, which is similar to the 300 HU used in this paper. The difference in our methods was that we defined any tissue in the model below the bone threshold to be bone marrow.

Changes in the text: We have modified our text to explain how the medullary cavity region was determined (Page 9, line 171). "The medullary cavity was defined using an HU cutoff for bone of 226 HU, corresponding to an elastic modulus of 2462 MPa. All regions in the model with a modulus lower than this were considered to be part of the medullary cavity; this was confirmed by visual inspection of these regions."

Comment 6: Line 169: Why are the yield loads mentioned in the Methods section instead of the Results section?

Reply 6: Thank you for pointing this out. We agree that the yield loads should be in the Results section.

Changes in the text: We have modified our text as advised and moved yield loads to the Results section (Page 11, line 216). "Yield loads for individual samples were 867 N (C1), 1028 N (C2), 157 N (B1), 403 N (B2), 471 N (B3), 7.7 N-m (T1), 5.0

N-m (T2), and 3.8 N-m (T3)."

Comment 7: Line 180: "It was assumed that yield strain of bone is similar between humans and canines." Is there any supporting literature?

Reply 7: Thank you for the comment. We were unable to find any recent literature reporting canine yield strain or the differences between canine and human yield strain. Because of this, we removed this assumption from the manuscript. However, this value of 0.0073 was used in previous finite element study of canine bone (Laurent et al. 2016), which is why we made this assumption.

Changes in the text: We modified the text to remove this assumption and support the yield strain value that we used (Page 11, line 198). "The threshold strain for this criterion was 0.0073, which is the yield strain of human bone in tension (10) and has been used in previous canine studies (8)."

Comment 8: Table 1: It would be useful to add the number of samples for each loading type to the table.

Reply 8: Thank you for pointing this out. We agree that this information should be included.

Changes in the text: We have modified our text as advised by adding the number of samples to Table 1 (Page 12, Table 1, line 228).

Table 1: Experimentally determined average values of yield load, stiffness, and maximum principal strain (measured at the strain gauge location when yield occurred).

Loading type	**Stiffness (N mm/rad for torsion, N/mm for bending and compression)**	**Maximum Principal Strain (με) at Osteotomy**	**Yield Load/ Torque (N m for torsion, N for bending and compression)**
Bending (N=3)	28.9 ± 13.9 N/mm	9502 ± 2018	343 ± 135 N
Torsion (N=3)	16.9 ± 8.5 N mm/rad	5197 ± 343	5.50 ± 1.61 N m
Compression (N=2)	421 ± 116 N/mm	5354 ± 652	947 ± 81 N

Comment 9: Table 2: The ranges are quite large. The authors conclude that the strain error can be determined by FEA with high accuracy (line 228), but also show a range of errors between -11 and 32%. For yield load the authors state they find moderate accuracy, but these results range between -49 and 65%, which is quite a large deviation in my opinion. What is the definition of high accuracy and moderate accuracy?

Reply 9: Thank you for the comment. We agree that these definitions of high and moderate accuracy may not be appropriate. We still believe these study results are acceptable, as the yield load errors were comparable to other published papers.

Changes in the text: We have modified our text as advised by comparing the stiffness, strain, and yield accuracies to each other rather than an absolute statement of “high accuracy” or “moderate accuracy” (Page 15, line 266). “Overall, the finite element model was most accurate in predicting stiffness, followed by strain, with yield load have the lowest accuracy.” We also added a comparison to other paper regarding accuracy in terms of R2 values and yield load, supporting the acceptance criteria used in this study (Page 16, line 311). “When performing a linear regression on the relationship between experimental and FE-predicted yield, R^2 values of 0.94 for compression/bending and 0.88 for torsion were observed. These values are similar to previous research that used CT-based FE models to assess fracture likelihood in metastatic femurs which reported R^2 values of 0.90 and 0.93 for intact bones and bone with lesions, respectively (12). Two separate groups utilized similar FE methods in femurs and reported R^2 correlations of 0.78 (22) and 0.88 (23). This suggests that the proposed modeling procedure possesses similar accuracy to other published methods.”

Comment 10: Figure 5: Why is there only an error bar for the experimental data, and why to the error bars represent 10% instead of a standard deviation/standard error/range?

Reply 10: The 10% on the experimental data was supposed to be a visual representation of what +- 10% error would look like graphically. Because each bar on the graph is a single data point and not an average, standard deviations could not be used. However, we agree that the 10% is confusing, so we removed the error bars from the graphs.

Changes in the text: We modified the text by removing error bars from the graphs (Pages 13-14, Figures 5-6).

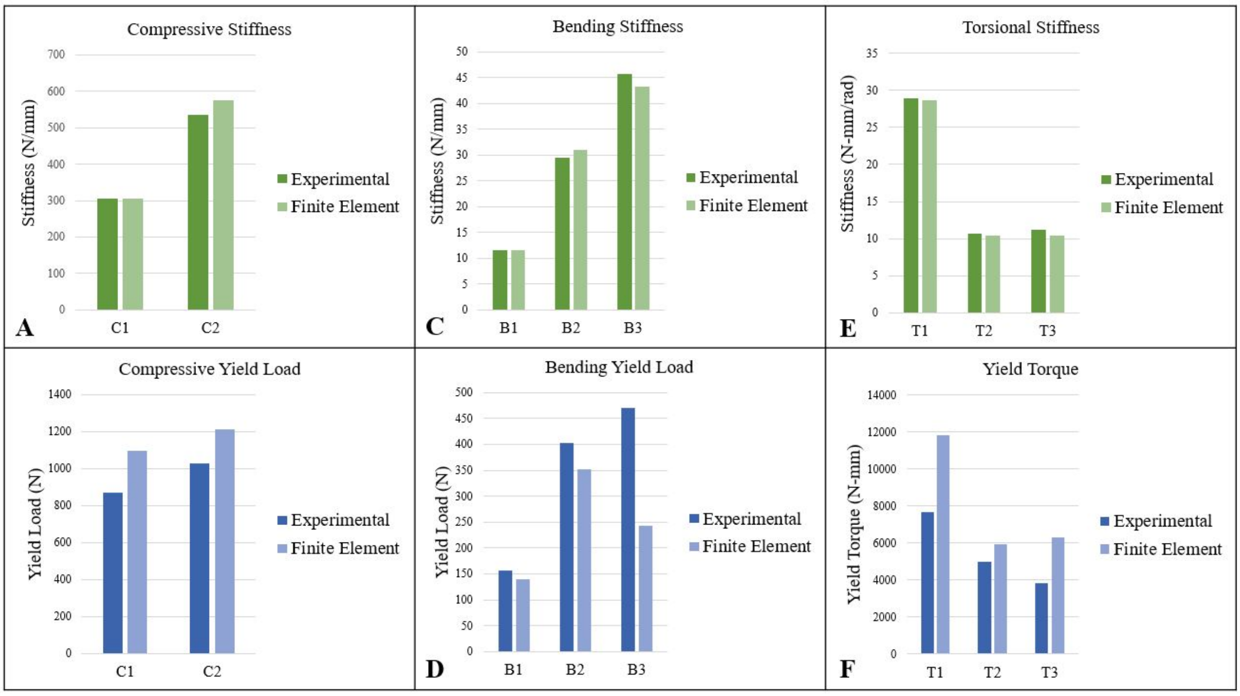


Figure 1: Comparison of experimental and finite element stiffness and yield load in compression (A, B), bending (C, D), and torsion (E, F).

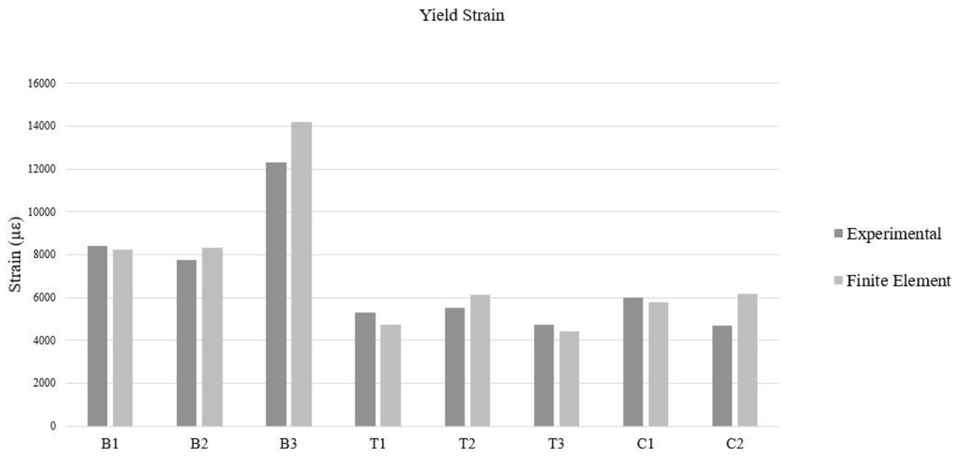


Figure 2: Comparison of yield strain between experimental and finite element data.

Comment 11: Line 225: The results of the three different loading conditions are analyzed with one t-test. Why do three loading conditions if you are going to analyze everything together? If you have three loading conditions, you also need to analyze them separately.

Reply 11: Thank you for bringing up this flaw in our analysis. We agree that loading conditions are more appropriately presented separately. We removed statistical analysis from the paper due to low sample size, eliminating the need for more t-tests. For the added linear regression analysis, we analyzed torsion separately from bending and compression. The bending and compression conditions were grouped together in this case due to the sample size of 2 for compression, which would have resulted in a misleading R^2 value of 1.0.

Changes in the text: We have modified our analysis procedure by separating torsion from bending and compression in the linear regression analysis (Page 14, line 248). “Plots of the regression are shown in Figure 7 for bending and compression (Figure 7A) and for torsion (Figure 7B).”

Comment 12: Line 232: What is ROI location?

Reply 12: Thank you for the question, we did not properly clarify this within the Results section. The ROI is at the location of the strain gauges, which is on the surface opposite the defect, as explained in Methods (see Page 6, line 123).

Changes in the text: We have modified our text as advised by clarifying the ROI location within the Results section (Page 15, line 267). “Some of the error in the strain-based yield load predictions likely resulted because the location of the fracture did not occur at the ROI (location of strain gauge measurements) in all samples, particularly for torsion samples”

Comment 13: Line 244: It should be mentioned that the fact that there was only a small number of samples and the fact that all loading conditions were analyzed together is a limitation of these results.

Reply 13: Thank you for the comment. We agree that the small sample size is a limitation, though we believe the size was adequate for the purposes of this preliminary study. The loading conditions have been separated out in the updated version of the manuscript to assist with interpretation of the data.

Changes in the text: We modified our text as advised by acknowledging small sample size as a limitation (Page 17, line 331). “The main limitation was that this study had a small sample size; additional samples are needed in future studies to support validation of the models more robustly.”

Comment 14: In the Discussion, it would be nice to add a comparison with the FE models of other groups (as mentioned before).

Reply 14: The authors agree that the addition of a comparison to other models would enhance this work.

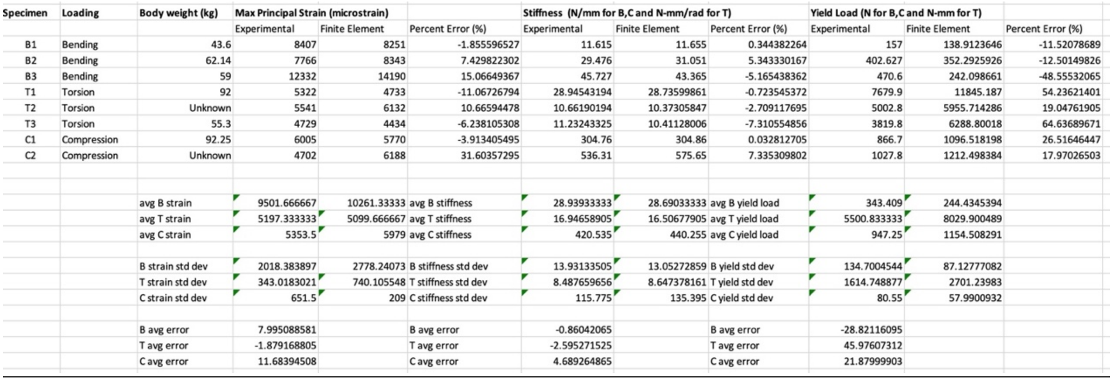
Changes in the text: We modified the text as advised to include comparisons to other papers regarding model accuracy and model setup (Page 16, line 311). “Results from this study were compared to similar fracture prediction studies. When performing a linear regression on the relationship between experimental and FE-predicted yield, R^2 values of 0.94 for compression/bending and 0.88 for torsion were observed. These values are similar to previous research that used CT-based FE models to assess fracture likelihood in metastatic femurs which reported R^2 values of 0.90 and 0.93 for intact bones and bone with lesions, respectively (12). Two separate groups utilized similar FE methods in femurs and reported R^2 correlations of 0.78 (22) and 0.88 (23). This suggests that the proposed modeling procedure possesses similar accuracy to other published methods.

The modeling procedure of this study mainly differed from other CT-based FE models in the definition of material property. Stadelmann et al. set a constant elastic modulus (10 GPa) for their publication on human vertebrae (13), while others used HU and density equations to calculate modulus without considering the effect of bone marrow on mechanics (24, 25). The technique outlined in this study used similar equations for bone (though for canine rather than human) but set bone marrow modulus to a constant value. When investigating methods of applying material properties, it was found that using these equations for the whole model (including the medullary cavity) had a 30% higher error than proposed method, while neglecting bone marrow had a 38% higher error. This suggests that bone marrow does affect bone mechanics and should be accounted for."

Comment 15: Line 258/supplementary material: To enable better interpretation of the results, it would be helpful to add the body weights of the dogs to the table in the supplementary material.

Reply 15: Thank you for the comment. We agree that body weights should be added.

Changes in the text: We added body weight data to the supplementary material (Page 23, line 456).



Specimen	Loading	Body weight (kg)	Max Principal Strain (microstrain)			Stiffness (N/mm for B,C and N-mm/rad for T)			Yield Load (N for B,C and N-mm for T)		
			Experimental	Finite Element	Percent Error (%)	Experimental	Finite Element	Percent Error (%)	Experimental	Finite Element	Percent Error (%)
B1	Bending	43.6	8407	8251	-1.855596527	11.615	11.655	0.344382264	157	138.9123646	-11.52078689
B2	Bending	62.14	7766	8343	7.429822302	29.476	31.051	5.343330167	402.627	352.2925926	-12.50149826
B3	Bending	59	12332	14190	15.06649367	45.727	43.365	-5.165438362	470.6	242.098661	-48.55532065
T1	Torsion	92	5322	4733	-11.06726794	28.94543194	28.73599861	-0.723545372	7679.9	11845.187	54.23621401
T2	Torsion	Unknown	5541	6132	10.66594478	10.66190194	10.37305847	-2.709117695	5002.8	5955.714286	19.04761905
T3	Torsion	55.3	4729	4434	-6.238105308	11.23243325	10.41128006	-7.310554856	3819.8	6288.80018	64.63689671
C1	Compression	92.25	6005	5770	-3.913405495	304.76	304.86	0.032812705	866.7	1096.518198	26.51646447
C2	Compression	Unknown	4702	6188	31.60357295	536.31	575.65	7.335309802	1027.8	1212.498384	17.97026503
		avg B strain	9501.666667	10261.33333	avg B stiffness	28.93933333	28.69033333	avg B yield load	343.409	244.4345394	
		avg T strain	5197.333333	5099.666667	avg T stiffness	16.94658905	16.50677905	avg T yield load	5500.833333	8029.900489	
		avg C strain	5353.5	5979	avg C stiffness	420.535	440.255	avg C yield load	947.25	1154.508291	
		B strain std dev	2018.383897	2778.24073	B stiffness std dev	13.93133505	13.05272859	B yield std dev	134.7004544	87.12777082	
		T strain std dev	343.0183021	740.105548	T stiffness std dev	8.487659656	8.647378161	T yield std dev	1614.748877	2701.23983	
		C strain std dev	651.5	209	C stiffness std dev	115.775	135.395	C yield std dev	80.55	57.9900932	
		B avg error	7.995088581		B avg error	-0.86042065		B avg error	-28.82116095		
		T avg error	-1.879168805		T avg error	-2.595271525		T avg error	45.97607312		
		C avg error	11.68394508		C avg error	4.689264865		C avg error	21.87999903		

Comment 16: Line 264: The limitations about the low number of samples and low power are missing.

Reply 16: Thank you for the comment. We agree that these are limitations that should be acknowledged.

Changes in the text: We modified our text as advised to acknowledge sample size and low power as limitations (Page 17, line 331). "The main limitation was that this study had a small sample size; additional samples are needed in future studies to support validation of the models more robustly."

Comment 17: As mentioned before, the conclusion is too strong considering the low sample size and the fact that canine radii were used instead of human material.

Reply 17: Thank you for the comment. We have modified the conclusion accordingly.

Changes in the text: We have modified our text as advised by rewording the conclusion (Page 18, line 355). "In conclusion, the finite element technique presented in this study shows promising accuracy in predicting bone fracture mechanics in canine radii with artificial osteosarcoma defects. Although more work is needed in order to translate to humans, this method could eventually provide clinicians with quantitative data to support decisions regarding surgical intervention for patients with osteosarcoma or bone metastases."

Reviewer B

Comment 1: This study used well established methods for creating simple linear elastic finite element model from long bone CT data. Simple loading is applied as per experimental conditions through testing fixtures (with no soft tissue forces). While the paper is sound it is not particularly novel. Dr Snyder's group has done similar work in human long bones using both FE models and structural rigidity measures to evaluate fracture risk (and compared this to MIRELS scoring) in a number of papers published over the past decade and longer. The authors have the Derikx paper in their references – "Some studies have also modeled metastatic lesions from human cadaveric tissue (12, 13)" – with no comparison to their methods or findings. The Snyder group also looked at fracture risk in benign skeletal neoplasms in children finding CT based structural rigidity measures worked very well in this scenario. As such, are FE models needed in these simplified scenarios?

Reply 1: Thank you for the comment. We chose to use simple loading conditions (compression, torsion, and bending) in order to better represent a wide range of loading; since complex loading involves some combination of these simple loads, if the model can correctly predict mechanics of simple loads, it should be able to predict complex loads as well. Our future work will aim to support this assumption. We used a linear elastic model to reduce complexity and computational time in the goal of clinical feasibility. We believe that FE models are still needed in these simplified scenarios; as noted in the Introduction, non-FE methods such as MIRELS scoring lack patient-specificity geometric features. Thus even simplified patient-specific models should provide improved fracture prediction due to greater accuracy in the geometric modeling than non-FE methods.

Changes in the text: We have modified our text as advised to include comparisons to the Derikx paper (Page 16, line 311). "When performing a linear regression on the relationship between experimental and FE-predicted yield, R^2 values of 0.94 for compression/bending and 0.88 for torsion were observed. These values are similar to previous research that used CT-based FE models to assess fracture likelihood in metastatic femurs which reported R^2 values of 0.90 and 0.93 for intact bones and bone with lesions, respectively (12)."

Comment 2: Did you consider structural rigidity measures?

Reply 2: Thank you for the comment. No, we did not consider structural rigidity in our models. We did, however, measure stiffness, which is similar to rigidity in that it considers resistance to deformation. Because of this, we felt that measuring rigidity would be redundant and would make the modeling procedure more time consuming. Furthermore, rigidity would incorporate specimen area, which is not constant along the length of the sample and would therefore need to be averaged across several points, reducing accuracy.

Changes in the text: No changes were made to the text.

Comment 3: Note the simulated model is not necessarily reflective of metastatic or primary tumors – a limitation that has been acknowledged in many studies using a drill hole to simulate pathology. It would be helpful to report R2 values to allow easier comparisons with earlier work.

Reply 3: Thank you for bringing up R2 values as an analysis technique; we agree that reporting R2 would help with comparisons. We also agree that our model is not yet representative of tumor tissue. In the future, we plan to expand this modeling procedure to bones with tumors now that it has been validated on healthy bone.

Changes in the text: We modified our text as advised to acknowledge the limitation of using a drill hole (Page 18, line 347). "Although the osteotomy simulates the weakening of the bone that occurs in active lesions, it does not take into account structural and material properties of the tumor. Future work will further investigate the use of FE models on limbs with clinically diagnosed osteosarcoma to better represent the effect of tumors on fracture mechanics." A linear regression with R2 values was also added for comparison (Page 14, Figure 7, line 251).

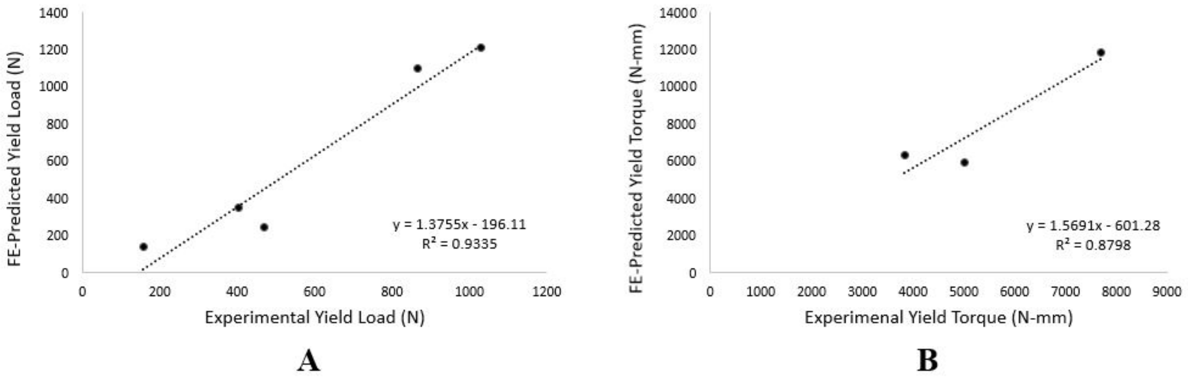


Figure 7: Linear regression comparing experimental and finite element yield for compression and bending (A) and torsion (B). Compression and bending were grouped together to ensure sample size was sufficient to perform linear regression.

Comment 4: As acknowledged in the limitations the time taken to currently construct the models is not feasible for clinical translation – as such the enthusiasm for the translational aspect of this work in tempered.

Reply 4: Thank you bringing up this concern. We agree that the time it takes is a limitation of the study. Before moving to clinical translation, it will be critical to automate portions of the process to make computational modeling more attractive for clinical use, as noted in the Discussion. Nevertheless, we feel that advances in computational modeling and computational power can assist the clinical community.

Changes in the text: No changes were made to the text.

Comment 5: Finally, CT scanning (vs 2D xray imaging) of such lesions may not be standard of care limiting development of specimen specific models using the approach described herein.

Reply 5: Thank you for pointing this out. We agree that this may not always be a standard of care for cancer patients, though we anticipate that it may become more common as the benefits of diagnostic computational modeling grow more widely known.

Changes in the text: We modified our text as advised to acknowledge this limitation (Page 17, line 338). “The need for CT scanning currently also limits clinical feasibility, as this isn’t always the standard of care for patients with bone tumors (plain radiographs are more commonly used). However, the prevalence of CT scanning in this population may increase as more diagnostic computational modeling technology becomes available.”