*Comment on the updated change score results:*

When reviewing the analysis for the results presented in Table 3, we noticed a mistake in the code used to generate these values. We have fixed this problem and have updated the Table 3 results.

*1) Tibial sesamoid data (from Reviewer 2) - is it too late to run these numbers from before?*

There was no variation in this outcome, so we were unable to perform this analysis.

2)*According to the authors "The primary purpose of this study was to compare clinical outcomes and complications at a minimum of 1 year postoperatively between the scarf osteotomy and the modified Lapidus procedure in patients indicated for operative correction of their hallux valgus deformity."  This is framed as a conventional pre-to-post type of analysis, but the authors have undertaken some unconventional approaches in the analysis.*

While the use of TMLE may be uncommon, it is not an incorrect approach. The pre-to-post design is also not relevant to the use of TMLE; TMLE is a general analytic framework that can be adapted to any study design.

3) *Lines 64-65:  The authors state they hypothesize that there would be no difference between the procedures in the outcome measures --- this implies they plan to use an equivalency analysis rather than a superiority trial.   However, everything I read in the Methods section shows that they powered the sample and analyzed the outcomes as per a superiority trial.  This needs to be clarified as sufficiently powering for an equivalency analysis requires far more subjects compared to a superiority trial.*

We agree that our current language implies that this is a study of equivalency. We have re-phrased to: “We tested the null hypothesis of no difference in clinical or radiographic outcomes between the two groups.”

4) *Lines 83-98:  Here the authors describe the a priori power analysis but there is no mention of powering based on equivalency between groups.  Also it is a bit odd to describe how the two groups compare by gender, age etc in the Methods rather than the Results section.*

We have rewritten so not to describe as an equivalency trial. We have also moved this to the Results section.

*5) Lines 141-160: Statistical analysis*

*Here the authors discuss using TMLE as an analytic method to evaluate what appears to be a conventional pre-post analysis.  To my understanding TMLE is employed for big data/data mining projects where building a causal model is the end product.  The does not appear to be the situation in this case so the authors need to defend their use of TMLE rather than conventional regression techniques that also can account for covariates.  Methods such as propensity score weighting and inverse probability weighting are commonly employed to do this.   TMLE is not necessarily appropriate for this relatively small dataset and for what seems to be a conventional pre -post framework.*

The use of TMLE is orthogonal to this study being a pre-post design. We also disagree that TMLE is only valid for “big data”. TMLE is a general analytic framework for building nonparametric, asymptotically linear estimators. While TMLE is generally used in the context of causal inference, it remains a valid analytic strategy when causal identification assumptions do not hold. TMLE has desirable statistical properties that make it superior to other analytic approaches such as conventional regression adjustment, and g-computation or IP weighting. Most notably, TMLE is doubly robust; it will remain a consistent estimator of the target estimand if either the outcome model or treatment model is correctly specified. If both the outcome model and the treatment model are correctly specified, then TMLE is asymptotically efficient. We have updated the statistical analysis section to highlight this, writing: “The CV-TMLE is a doubly robust estimator that remains consistent to model misspecification of at most either the outcome regression or the propensity model.” TMLE is also a substitution estimator while IPW is an estimating-equation based estimator. As such, IPW ignores global constraints of the statistical model and can produce probability estimates outside of 0 and 1, or in the case of PROMIS scores, produce estimates outside of 20 and 80. This is not the case with TMLE. Furthermore, TMLE can be used with machine learning algorithms while still yielding valid standard errors while the use of the other mentioned methods requires parametric models to remain root-n consistent. We have updated our results to use CV-TMLE instead of TMLE. CV-TMLE uses an additional layer of cross-validation, after the cross-validation used to establish the super learner weights, that protects against overfitting of the nuisance parameters by estimating the nuisance parameters out-of-sample.

Rose S., van der Laan M.J. (2011) Why TMLE?. In: Targeted Learning. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9782-1\_6

*6) Also in this paragraph lines 153-160: How exactly where these tests run as equivalency tests?  That normally employs establishing 2 borders (high and low) around the group differences and determining if the group differences exceed either one of those boundaries.  Please elaborate.*

This is not an equivalency study.

7) *More on TMLE:  For the family of machine learning techniques there are usually training and validation processes involved with the model development, yet there is no discussion or presentation of the results of those for this small database.  There is a reason why methods such as TMLE are appropriate for "big data" analyses and not for these small cohort studies -- to develop a model, especially a causal model for which these methods are intended for, requires the training and validation of big samples to develop confidence in the cross-validation.  Please explain how that confidence was developed in this sample and present the results of those training and validation runs*

TMLE is not a machine learning method but is a general analytic framework. As previously explained, TMLE has strong finite sample performance and is not only appropriate for “big data.” The machine learning aspect of this study was the use of the super learner algorithm for estimating the nuisance parameters that are then used with TMLE. The super learner was used with cross-validation. Indeed, the way the super learner is constructed is by building a stack of predictive algorithms that minimizes the cross-validated risk. Asymptotically, the super learner will outperform any of the individual algorithms included in its construction. The super learner has also been shown to demonstrate good finite sample performance in sample sizes as low as n = 100 (see Polley and van der Laan, 2010). We use the super learner because it is impossible to know what algorithm will perform the best in a certain sample. We did include a GLM in the super learner library, if a GLM is sufficient for this data then the super learner will assign it the most weight. We have now included a table that shows the weights applied to the individual prediction algorithms in the super learner library that minimizes the cross-validated risk. As previously mentioned, we have also updated our methodology to use CV-TMLE instead of just TMLE which introduces an extra layer of cross-validation to generate out-of-sample estimates of the nuisance parameters. This protects against over-fitting of the nuisance parameters which could affect the variance of the TMLE estimands.

Polley, Eric C, and Mark J van der Laan. 2010. “Super Learner in Prediction.” *Bepress*. bepress.

8) *Discussion section:  Every time the authors state the TMLE controlled for confounds would also apply to other regression methods -- it is not clear why TMLE was chosen for this analysis.  Also, it would be nice to see how controlling for the surgeon affects the results -- it truly is a major limitation of a study like this.*

Please see our previous responses for why we used TMLE over other methodologies. Because most surgeons only perform one procedure, adjusting for the surgeon would induce practical positivity violations and propensity score estimates would become unstable.