WINTER 2019/20 TERM 2 EPSE 592: Take-home Final Exam
Due date: Apr. 30th

- You may choose to write your answers by hand or using a computer. In either case, **please email me a PDF of your completed exam by the above deadline(s)**. I will then grade and make comments directly on your submitted PDF, and return an annotated copy to you.
- Please make sure you write your answers to these questions in your own words. Even if you work with a group to formulate your responses, do not just copy someone else's sentences/words.

CASE STUDY: "Stronger back muscles reduce the incidence of vertebral fractures: a prospective 10 year follow-up of postmenopausal women," Sinaki *et al.*, *Bone*, 30(6), 836–841 (2002).

Your task is to critique the study by Sinaki *et al.* (2002) published in *Bone*, a tier 1 research journal in Physiology, Histology, and Endocrinology. None of you are experts in this medical domain (and neither am I), but remember: we can still critique the methodology and inferences using the basic principles we have learned about statistical best practice and proper inferential reasoning. It is not necessary to fully understand all of the domain-specific jargon.

I strongly recommend that you read the study once and make brief notes about any items that grab your attention, both in terms of methodology employed and conclusions drawn. Then step away from the study for a day or two. Read the questions I have prepared below to help guide your critique, and then reread the study with an eye towards these specific questions. Make more notes. Then step away for another day or two, return, and reread the study a third time, this time making comprehensive notes to address my prompts.

Remember that sometimes you may not be able to answer all of these questions definitively. You only have the information that the study's authors communicate in their paper. But if you *cannot* answer a question definitively, then that can have repercussions for your assessment of the study's methodology and conclusions.

Finally, yes, there are a lot of questions here! But remember that many of these can be answered quite succinctly.

Question 1: Study design

(a) Would you classify this study as "exploratory" or as "confirmatory", and why? That is, are the researchers trying to generate plausible hypotheses to formally test in a follow-up study, or are they trying to confirm hypotheses that have been proposed?

(b) Comment on the sampling procedure used to generate the units of observation (patients). Was any random sampling performed? If sample units were excluded for any reason, do these reasons seem like good ones?

(c) Describe the target population of inference. That is, for what population is the study sample representative? It is this population to which all subsequent inferences can apply.

(d) What is the treatment (intervention) of interest? Does the study contain a control group? If so, how was "control" defined? Were patients randomly assigned to treatment or control? Was this randomization *restricted* at all?

(e) Were there any potential issues with patients not adhering to their prescribed treatment? Were measures taken - clinically or statistically - to try to adjust for these possible issues?

(f) What are the main response variables of interest? That is, what response variables are of clinical importance to the researchers?

(g) There were two phases of the study: an experimental treatment phase, and then a non-intervention follow-up phase. Generically speaking, what kind of research question could data from the first phase answer? What kind of research question could data from the second phase answer?

(h) Not all patients who completed phase one of the study also completed phase two. The authors chose to only analyze data on patients that completed both phases of the study. Why might this be a good idea from a methodological point of view? On the other hand, why could this be a bad idea? Given the main goals of their study, do you think this choice to only analyze data from patients who completed both phases of the study was a good idea?

Question 2: Statistical analysis and results

(a) Overall, how effectively do the authors communicate the results of their analysis? Remember: there are four items that should always be reported whenever some kind of statistical test is performed: a measure of *observed effect size*, a measure of *uncertainty* for that observed effect size, something indicating the type of statistical test performed, and a subsequent p-value.

(b) The main omnibus tests that the authors ran were (mostly) several repeated measures ANCOVAs. For the analysis on patient height (outcome variable) only: (i) which factor(s) were within-subject? (ii) Which factor(s) were between-subject? (iii) Which baseline covariate(s) were adjusted for? Is any indication given that the variety of RM-ANCOVA assumptions were checked for validity?

(c) On page 838, the authors describe a kind of rough power analysis. How useful is this information? Think about what can affect power and how the authors justify (or fail to justify) their power calculations. Does it seem like the power analysis did a good job of targeting believable effect sizes? That is, were the observed effect sizes comparable to the targeted effect sizes? Note: you can approximate standard deviations for outcome measures like 'height' by looking at the sample standard deviations in Table 1. [Bonus: there is also a critical *mistake* in their power analysis for the chi-squared test on the number of vertrebral bodies; can you find it?]

(d) Rerun the χ^2-test the authors performed on the "incidence of vertebral compression fracture," remembering to use a *continuity correction* (p. 838). Did the authors use a continuity correction? Would using or not using a continuity correction make a difference here under the significance threshold interpretation of p-values?

(e) Comment on the observed outcomes for patient *grip strength* (make sure to consider Figure 3B).

(f) Consider the reported outcomes on *bone mineral density* (BMD). Look at Table 1 to obtain the average BMD for each patient group at each of the 3 study time points (compare with Figure 4), then use these figures to calculate the sample difference between experimental and control group average BMD at each of the 3 study time points. Do the p-values reported in the text and caption of Figure 4 that attempt to test if these mean differences are significantly different from zero make sense when looking at the actual observed mean differences? Why or why not? What could explain the reason for any ostensible contradictions here?

(g) Consider the *physical activity score* outcomes. The BE experimental group reported higher physical activity scores in all categories of physical activity at 10 year follow-up. Why is this result important? Notice: the previous analyses on BMD and muscle strength did *not* account for differences in physical activity scores between treatment and control group.

(h) How big of an issue is an inflated type I error rate due to multiple comparisons here? To answer this, give an **approximate** count for how many *disjoint* statistical tests are being performed in this study. That is, try to count how many different AN(C)OVAs, RM-AN(C)OVAs, χ^2, Kruskall-Wallis, and *t*-tests have been performed. Remember: some of those t-tests are post hoc comparisons from an omnibus AN(C)OVA or RM-AN(C)OVA model; such post hoc tests are *not* to be counted as they are already adjusted for under the Tukey procedure, as stated in the text. Also note: your count does not have to be exact.

(i) Using your approximate count from the previous part for the total number of disjoint statistical tests performed, use a Bonferonni correction to adjust the original $\alpha = 0.05$ significance level. Which substantive results that were previously flagged as significant are no longer significant after this adjustment for multiple comparisons?

Question 3: Discussion and conclusions

(a) The authors report that the "incidence of hip fracture in this geographic location for this age group has been reported to be 3.85/1000 per year." Yet no one in the study sample sustained a hip fracture over the course of the study. Given the number of years over which the study was conducted and the number of patients in the sample, about how many hip fractures would we have expected to see using the 3.85/1000 figure? What are some reasons that could explain why our study sample saw *zero* hip fractures instead?

(b) Using the previous knowledge discussed in the study, assess the *prior believability* of the substantive hypotheses the authors claim evidence for. How should this change, if at all, your interpretation of the study's conclusions?

(c) The authors conclude: "benefits from participation in a 2 year back exercise course continued even 8 years after cessation." How convincing do you find this claim based on the evidence presented in the study? What benefits (if any) does the study provide weak or strong evidence for?

(d) The main conclusion of the study is contained in its title: "stronger back muscles reduce the incidence of vertebral fractures [in] postmenopausal women." How strongly do you agree with this conclusion? What kinds of caveats (if any) should be communicated to temper the generality of this conclusion? Do you think the authors do a sufficient job of communicating these limitations?

(e) Are there any other comments or critiques about the study or its conclusions you would like to add?