Data Analysis Project 2
Robert Han and Alyssa Peeples

Determine if you could repeat the analysis that is most essential to conclusions of the paper

The paper by Petrashek, Ye, and Buck describes a longevity effect of dosing C. elegans with a human antidepressant (mianserin). The paper itself is a followup study to an earlier paper (Melov et al. 2000) exploring the extension of life-span with superoxide dismutase/catalase mimetics.

Figure 1.c provides the rationale for the title of the paper. In it we see that mianserin at around 50 μM increases the average lifespan of *C.elegans* by more than 50%. The plotted means are compared with several mutants to provide context as to what pathways the drug is interacting in. Figure 1.b also shows this conclusion; however, we believe 1.c does a better job of showing this conclusion visually as it is more concise and shows the mean increase in lifespan rather than fraction of worms alive.

The original data was not readily available which is not surprising given that the paper was published in 2007 and the push for data transparency is more recent. That being said, we can infer what the methods for testing were as well as the data analysis using the figure and figure caption. The testing methodology likely involved raising several lines of *C. elegans* and measuring their lifespan when they are treated with varying concentrations of mianserin. As far as data analysis we can check the Supplementary Information to find the summary statistics used to generate the figure.

Using the information in Supplementary Table 2 of the Supplementary Information and inferences made about the methods, we can imagine what the original dataset might have looked like. There were likely four columns for: *C. elegans* strain, drug, drug concentration, and how many days the individual lived. Using that data the authors could have easily calculated the mean lifespan and standard error of the mean for each combination of strain, drug, and drug concentration which are the data points visualized in figure 1.c. Several other calculations for p-value, percent change, and number of individuals tested are made but not visualized in the figure. When actually plotting the data the authors chose to compare each mutant with the *N2* strain.

If we were to attempt to repeat the analysis we'd of course need the original dataset or to gather our own data. It may be useful to try to gather our own data since follow up studies appear to have gotten the opposite result (Zarse and Ristow 2008). Regardless, we can use R and the data analysis/visualization skills we've learned to replicate this result. Using the dplyr package we could subset the dataset into smaller ones based on strain. Then use native R functions to perform the statistical calculations (logrank test) to obtain the p-values. We actually don't need to use these calculations to generate the figure as R has a built-in `plotmeans()` function which can do the same. Otherwise, we'd call ggplot(strain) with: `geom_point(aes(x = concentration, y = mean_lifespan)`. We'd also need to include: `geom_errorbar(ymin = mean_lifespan - SEM, ymax = mean_lifespan + SEM)` to add the error bars seen in figure 1.c. Finally, figure 1.c has a horizontal line at 24 (days) which can be generated using: `geom_line(aes(y = 24, group = 1)`.

In summary, we believe that with access to a dataset, the most essential analysis done in the paper can be replicated using R with similar or varying results.

For each Questionable Research Practice discussed in Fraser *et al*:

The authors of the paper introduce the idea that of more than 80,000 drugs tested less than 100 were found to significantly increase the lifespan of C. elegans. The paper itself is focused on one of these drugs (mianserin) which showed the greatest increase in lifespan and postulates the avenues of the drugs action using several strains of C. elegans with serotonin signalling mutations. The supplementary information appears to show all relevant summary statistics for the drug-strain combinations for each figure so we might assume that questionable research practices like cherry picking and p-hacking have not occured; however without the raw data we need to be cautious in our evaluation.

I was concerned when I looked into the details of the logrank test and saw it is appropriately used on “censored data.” In this case, “censored” is a method of accounting for individuals who haven’t died yet at the time of data analysis. In the context of survival and lifespan, we can’t have data points with no “end” so to speak, so the “end” is not really the end but used to create an observation for data analysis. Due to the nature of censored data, we might raise the question of p-hacking by arbitrarily choosing a stop point for data collection; however we do not believe the censoring used in this analysis had a major impact in the conclusions drawn as many if not all of the conclusions can be drawn from the data before censoring occurs.

The methods section mentions that wells containing more than 19 animals were excluded from statistical analysis with no other explanation for the decision. While it could be innocuous, we do not have the data to check and so this may represent one example of data cherry picking. The decision to omit those data points falls under the umbrella of “researcher degrees of freedom,” and may be absolved by complete reporting and preregistration of the decision (Fortmeister, Wagenmakers and Parker 2017).

In the supplementary information we observe evidence of multiple testing for several experiments where the number of animals tested exceeds 19 (all cases). With no mention of a Bonferroni correction in the paper, the significance of increases in fraction alive (figure 1.b and figure 3.c) could be called into question. The approaches to this problem are to: report all tests, perform Bonferroni correction, and emphasize the preliminary nature of the finding (Fortmeister, Wagenmakers and Parker 2017).

I do not believe HARK-ing has occurred in this study. HARK-ing is the practice of generating a hypothesis after data has already been gathered. The authors appear to have started with a goal of finding specific drugs which extend the lifespan of C. elegans. After finding several with significant effects, they performed a detailed analysis to examine the mechanism by which a specific compound acts to extend longevity. The hypothesis of the paper appears to be that mianserin enhances C. elegans longevity by altering serotonin signalling. Justification for this hypothesis is well introduced and explained by the authors of the paper, and the summary data resulting from that hypothesis and subsequent testing is documented in the publication.

Design a follow-up study to replicate the most crucial experiment

Acknowledgements:

Robert's contribution: Determine whether the most important finding can be repeated, QRP discussion, and NR reporting summary checklist.

Alyssa's contribution: Follow-up study design.

Fallon and Brenna helped Rob come to the conclusion that the data cannot be found online and also shared resources from Professor Field that were relevant to our assignment.

Works cited

Forstmeier, Wolfgang, Eric-Jan Wagenmakers, and Timothy H. Parker. 2017. “Detecting and Avoiding Likely False-Positive Findings – a Practical Guide.” *Biological Reviews* 92 (4): 1941–68.

Melov, S., J. Ravenscroft, S. Malik, M. S. Gill, D. W. Walker, P. E. Clayton, D. C. Wallace, B. Malfroy, S. R. Doctrow, and G. J. Lithgow. 2000. “Extension of Life-Span with Superoxide Dismutase/Catalase Mimetics.” *Science (New York, N.Y.)* 289 (5484): 1567–69.

Petrascheck, Michael, Xiaolan Ye, and Linda B. Buck. 2007. “An Antidepressant That Extends Lifespan in Adult Caenorhabditis Elegans.” *Nature* 450 (7169): 553–56.

Zarse, Kim, and Michael Ristow. 2008. “Antidepressants of the Serotonin-Antagonist Type Increase Body Fat and Decrease Lifespan of Adult Caenorhabditis Elegans.” *PLOS ONE* 3 (12): e4062.