**Master of Biomedical Informatics Capstone Project Progress Report 2**

**Student Name: Yumin Wang**

**Student Email: yumin\_wang@hms.harvard.edu**

**Title of Project:** **The application of specification curve analysis in controversial questions in nutrition** (subject to change)

**Progress Report #: 2**

**Date Submitted: September 27, 2022**

**Mentor Name: Chirag Patel**

**Mentor Email: Chirag\_Patel@hms.harvard.edu**

**Co-Mentor: Dena Zeraatkar**

**Co-Mentor Email: dena.zera@gmail.com**

**Is IRB Approval Required:**  Yes  No

**Date of IRB Application: Not applicable**

**IRB Number: Not applicable**

**Work accomplished to-date (200-300 words):**

*Outline all progress completed to-date, detailing any issues that have been encountered with data collection, analysis, writing and IRB approval. Please differentiate clearly between work completed yourself versus work performed by your project mentor and/or your project mentor’s lab members.*

*My capstone project continued during the summer months of 2022. Based on the systematic review called “Red and Processed Meat Consumption and Risk for All-Cause Mortality and Cardiometabolic Outcomes”1. I identified 20 papers associated with red meat and all-cause mortality. For each paper, I noted article name, author name, study cohort size, inclusion-exclusion criteria, the definition of unprocessed and processed meat, definition of mortality, variables collected, adjusting variables, and analytical specifications and then combined and summarized each information and put them into a word document for guiding my research. Based on this word document, I created an excel document that counts the number of times each variable appears in these 20 papers, as well as finds the corresponding variables in NHANES2 dataset. We then decided to use NHANES 2007-2014 data and link NHANES2 data with two other databases: National Death Index (NDI)3 database to get participants’ time to mortality data, and USDA Food Patterns Equivalent Database (FPED)4 to get participants’ food intake data. By using R, I combined NHANES2, NDI3, and FPED4 to get a raw dataset with all variables of interest which include demographic, mortality, dietary intake, and other related variables. Then, I read each variable’s documentation and started the data cleaning process which involves: identifying missing data, creating new variables based on related variables, grouping variables into predefined categories, converting units, rearranging columns, formatting variables, etc. After doing these, we got an analytical dataset where we would start our analysis. First, I exclude participants based on exclusion criteria summarized from literature and got 10,000+ complete cases with 1,000+ all-cause mortality. Then, I read a paper called “Specification curve analysis”6 by Uri et al. and “specr” package5 in r to learn the methodology and coding behind specification curve analysis. Finally, by using our analytical dataset, I run specification curve analysis using cox proportional hazard regression model7, setting unprocessed red meat as the exposure variable and time to mortality as the outcome variable, and get effect estimates, 95% confidence interval, p-value for hazard ratio, for 1000+ various model specifications. We are now in the process of interpreting the results and we are going to start to write our manuscript soon. Our project does not involve IRB approval. And up until now, all challenges have been solved in some way. I completed all the work listed here on my own with guidance from my mentor and co-mentor.*

**Modifications to original project plan (Bulleted List):**

*If changes have been made/need to be made to your project plan since the time of your project proposal based on oversight feedback, consultation with your project mentor, and/or any difficulties mentioned above, please detail these changes.*

*(1) We will not be doing alternative specification curve analysis for saturated fat intake and cardiovascular disease because time is tight and we think we have enough material on our hands to convey our message.*

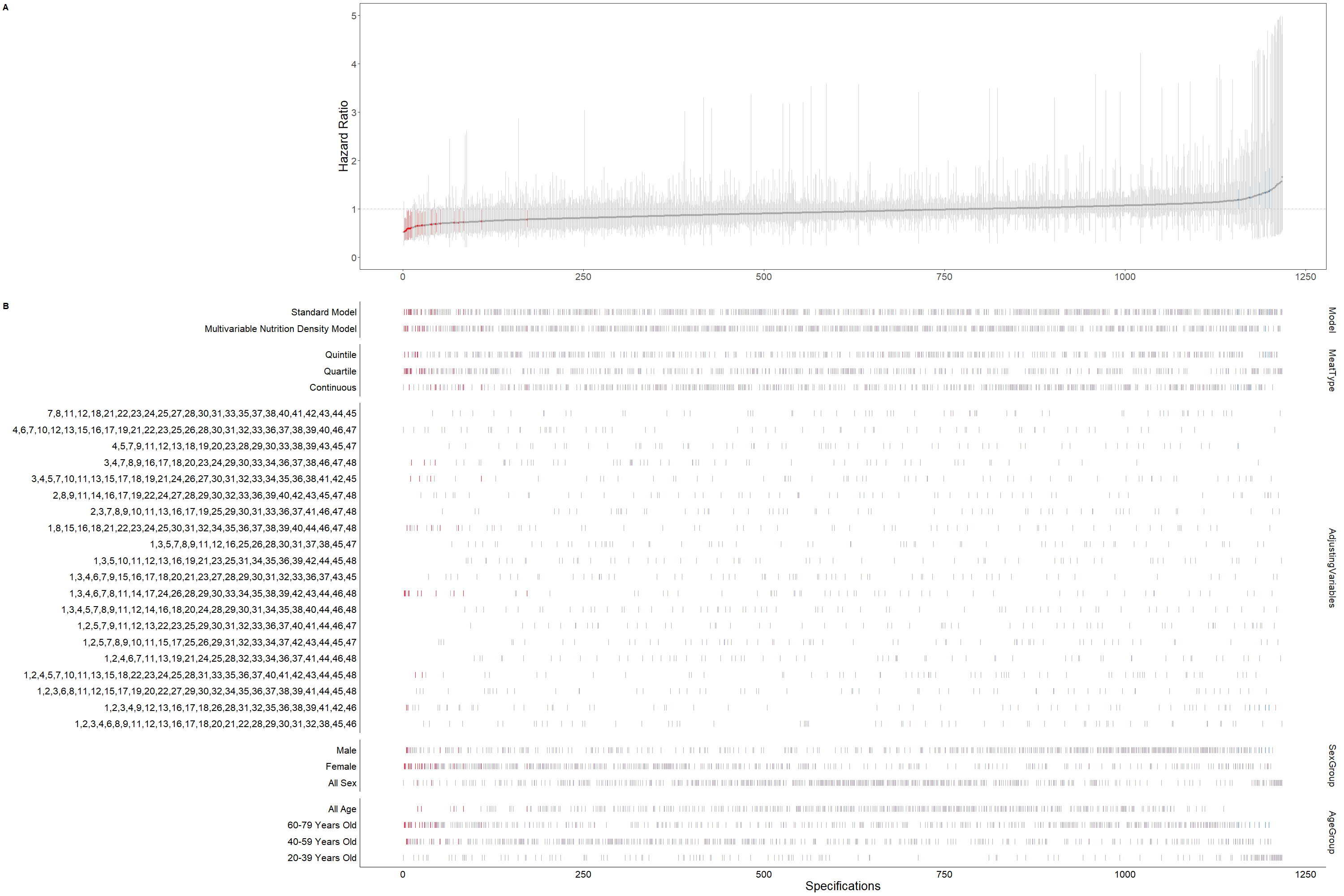
*(2) We will not statistically test how inconsistent the associations are with the null hypothesis by combining the results from all specifications because our main focus is to convey the message of vibration of effect estimates exist in nutritional research.*

**Does the work accomplished so far align with your project proposal timeline?**

Yes  No

**Results**

Either below or attached, include any and all results obtained so far in the course of your capstone project.



This graph is the main result we obtained so far. We examined the hazard ratio for all-cause mortality associated with the consumption of unprocessed red meat. A total of 1000+ model specifications are included in this graph. These specifications are obtained through the combination of the following individual model construct: 2 nutrition models (standard residual model and nutrition density model), 3 types of unprocessed red meat variable (quintiles, quartiles, continuous), 20 combinations of 48 adjusting variables (number denotes index, and an index corresponds to a specific variable), 3 sex groups (male, female, all sex) and 4 age categories (20-39, 40-59, 60-79, all ages). The blue dot in the upper graph depicts one significant and HR>1 effect by one specification. The red dot in the upper graph depicts one significant and HR<1 effect by one specification. The grey dot in the upper graph depicts non-significant results. The bar in the upper graph depicts a 95% confidence interval for one HR effect estimates by one specification. The bars vertically align below in the lower graph depict the analytical specification used for the corresponding hazard ratio effect estimate in the upper graph. The standard model is defined as treating unprocessed red meat as grams and adjusted for total energy. The multivariable nutrition density model is defined as treating unprocessed red meat as grams/kcal and adjusted for total energy. If quintile or quartile is used, then HR is defined as comparing the highest quartile (quintile) versus the lowest quartile (quintile). If continuous unprocessed red meat is used, then HR is defined as 100 grams increase for the standard model and 100 grams/2000kcal increase for the density model.

We can see the vibration of effect using different model specifications: some HR are above 1, and some HR are below 1, some significant results are above 1, and some significant results are below 1, and under which circumstances we can get different conclusions and that may explain the controversy in nutritional literature, especially in terms of unprocessed red meat and all-cause mortality. We may conclude that specification curve is a better way to examine the true relationship between a nutritional exposure and an outcome.

**Next Steps (Bulleted List):**  *List out the steps you plan to complete before your next submission.*

*(1) Discuss and interpret the graph and make more graphs if necessary. Maybe consider processed meat. Or maybe consider more specifications.*

*(2) Find a way to beautifully display these specification graphs.*

*(3) Draw other two graphs: (1) Baseline characteristics graph (2) Flow chart of inclusion-exclusion*

*(4) Start to write the manuscript, including the introduction, method, results, conclusion, abstract, etc.*

*(5) Revise the manuscript before submission on Nov 14th*

*(6) Prepare a presentation*

*(7) Potentially publish the paper*

**Journal Selected for Submission:**

Our aim is to publish this paper, so we are going to write our manuscript first and then decide exactly which journal to submit first.

Our potential journals for publication are:

1. The American journal of clinical nutrition

2. Journal of clinical epidemiology

3. BMJ medicine

4. BMJ

5. Annals of Internal Medicine

6. European journal of clinical investigation

We selected these journals based on previous publication of method related studies, and also based on method research within the journal’s scope.

**Journal Instructions for Authors Link:**

Here we give two Journal Instructions for Authors Link for two most likely submission journals:

The American Journal of Clinical Nutrition: https://academic.oup.com/ajcn/pages/General\_Instructions

Journal of clinical epidemiology: https://www.jclinepi.com/content/authorinfo

**References:**

[1] Zeraatkar, Dena, et al. "Red and processed meat consumption and risk for all-cause mortality and cardiometabolic outcomes: a systematic review and meta-analysis of cohort studies." Annals of internal medicine 171.10 (2019): 703-710.

[2] <https://www.cdc.gov/nchs/nhanes/index.htm>

[3] https://www.cdc.gov/nchs/ndi/index.htm

[4] https://data.nal.usda.gov/dataset/food-patterns-equivalents-database-fped

[5] <https://cran.r-project.org/web/packages/specr/vignettes/specr.html>

[6] Simonsohn, U., Simmons, J.P. and Nelson, L.D., 2020. Specification curve analysis. Nature Human Behaviour, 4(11), pp.1208-1214.

[7] Cox, David R. "Regression models and life‐tables." Journal of the Royal Statistical Society: Series B (Methodological) 34.2 (1972): 187-202.