Questions for June 29 meeting:

1. Definition of red meat and processed meat

Red meat: mammalian meat

Processed meat: white or red meat preserved by smoking, curing, or adding chemical compounds (looks like white meat is included)

My question is, because the systematic review conducted separate analyses for **unprocessed red meat**, processed meat, and mixed unprocessed red and processed meat, which are we focusing for our research?

1. Outcome of interest: all-cause mortality, cardiovascular mortality, cardiovascular disease, stroke, MI, type 2 diabetes, anemia, quality of life, satisfaction with diet.

Question: Are we just focusing on **all-cause mortality**?

1. Can you give me a template which include what kind of information I should pay attention to when reading the papers? Like a very general point by point guidelines like formula form or adjusting variables**? How much detail I should pay attention to?**
2. What’s our next steps?

Summarize all the information, find variables in NHANES.

Questions for July 20th meetings**:**

**The NHANES I clicked is 2009-2010, because mortality data linked from CDC is up until 31 December 2019, we have 9 years of follow up.**

21. Could not find time to event information? I mean 0, 1 death event and follow-up time for each person in NHANES.

Solved: https://www.cdc.gov/nchs/data-linkage/mortality-public.htm

23. Does accident constitute all-cause mortality?

YES

24. How many participants and death are we going to examine? The population size from literature examining this relationship using NHANES range from 9000-20000. For example, for NHANES 2009-2010, there are 10537 participants, among them 4027 are ineligible because age<18, leaving a total of 6510 eligible, and among them 861 death occurred. If these participants and death is not enough, then should we consider 2009-2010 NHANES and 2010-2011 NHANES and more NHANES combined?

**2000-2015. stay to the same format. 65000, 8000 death.**

1. Which age group are we focusing on? For example, if we are focusing on children 6-19, then education variables DMDEDUC3 will be used, if we are focusing on adults 20+, then DMDEDUC2 will be used. This also determines marital status because children don’t marry and also determine other things that children do not do. Also, mortality data for less than 18 is ineligible.

Adults

2. For income, what kind of income do we use? Annual household income or Annual family income, or monthly?

Annual family

3. For systolic blood pressure, which reading are we using? 1st, 2nd, 3rd, 4th?

Use 1st one and replace 2nd or 3rd or 4th if missing.

4. Can we discuss to determine the variables for alcohol drinking? I used Avg # alcoholic drinks/day -past 12 months

Use this variable.

5. Can we discuss to determine the variables for smoking?

https://shouldiscreen.com/English/pack-year-calculator

6. Can we discuss to determine occupation variables?

OCD241 - Occupation group code: current job

7. Can we discuss to determine history of cardiovascular disease? I only used coronary heart disease but I think this is only part of cardiovascular disease.

Combine coronary heart disease and stroke

8. Can we discuss to determine history of depression in NHANES? They just asked about lower mood…..

PQ020 - Feeling down, depressed, or hopeless

9. There are two dietary 24-h recalls in NHANES, which one do we use? Day 1 or Day 2?

Use the average. Combine Day 1 and Day 2.

10. Do you already have code for manipulating NHANES data, I mean merging combining dataset together, and most importantly, any analytical requirement that need to be paid attention to, like is sample weights needed to be included in cox regression? need a template code for this? Which software do we use? Need more information on this? SAS or R? I prefer R.

**Chirag have some code for this. We don’t need sample weights. We prefer R.**

11. It looks like we need to link food codes to individual food and create variables such as vegetables intake (vegetables intake), fruit intake, red meat intake (red meat includes many sources) so we need to determine on the exact definition of these things. The problem here is that food is mixed.

Consider variation in definition. Dena will find how to define and combine food in literature.

12. There are so many eggs? Which variable do we need to combine? Need help with this. Exact definition.

13. It looks like we need to link prescription medications with drug information to help find drug use.

We do this.

14. Could not decide menopausal status (premenopausal, postmenopausal)? Only have information about Age at last menstrual period

RHQ060 - Age at last menstrual period 1 year after current age then postmenopausal.

15. Could not decide hormone therapy users? In literature: in postmenopausal women (not taking hormone therapy, taking hormone therapy. But in NHANES, {Have you/Has SP} ever used female hormones such as estrogen and progesterone? Please include any forms of female hormones, such as pills, cream, patch, and injectables, but do not include birth control methods or use for infertility.

This is good.

16. For socioeconomic status, how to determine poor, middle income, high criteria using PIR.

We discuss later.

17. Wealth score is created by Using multiple correspondence analysis, the wealth score was created based on occupation, house ownership, house structure, house size, having a bath in the residence, as well as a personal car, motorbike, black/white TV, color TV, refrigerator, freezer, vacuum, and washing machine. Can we remove wealth score? Only in one literature.

Delete it.

18. NO find for Consumption of trans-fat, family history of cancer, could not find rural or urban residence?

Dena will find trans-fat, we don’t cancer information, we don’t need urban resident.

19. In an effort to control for severity of illness, we included self-reported health status as well as comorbid conditions. A variety of conditions were assessed in the NHANES II. Comorbidities were positive responses in the baseline interview to questions regarding whether a physician ever told the patient that he or she had each of the following conditions: cirrhosis, diabetes, high blood pressure, heart failure, heart attack, stroke, hardening of the arteries, rheumatic fever, rheumatic heart disease, heart murmur, ulcer (peptic, stomach, duodenal), chronic enteritis, ulcerative colitis, spastic colon or mucous colitis, gallstones, hepatitis, yellow jaundice, pleurisy, low blood pressure, cataracts, glaucoma, thyroid disease, polio or paralysis, goiter, hiatus hernia of the diaphragm, cancer, benign tumor, trouble with blood not clotting properly, loss of blood from stomach or bowels, nervous breakdown, neck injury, back injury, anemia, arthritis, gout, asthma, chronic bronchitis, emphysema, tuberculosis, and kidney problems. The Charlson Comorbidity Index was calculated from the responses to these questions. **This is very complex; can we remove this comorbidity index?**

**Remove**

20. Years of entering cohort is based on year of NHANES?

Yes.2009-2010 record it as 2009.

22. How we define unprocessed red meat?

Questions for July 27th meeting:

(1) **Smoking:** Could not decide the variable for **smoking:** only have

Age started smoking cigarettes regularly

Age last smoked cigarettes regularly

Avg # cigarettes/day during past 30 days

During the past 5 days, on the days {you/he/she} smoked, how many cigarettes did {you/he/she} smoke each day?

<https://shouldiscreen.com/English/pack-year-calculator>

Combine SMQ040, SMQ050Q, SMQSMD650

Current non-smoker, current smoker with categories, people who quit smoking

(2) **Occupation:** NHANES 2015-2016 does not have OCD241 - Occupation group code: current job

Remove 2015-2016

(3) **Sleep:** NHANES Sleep hours format for 2015-2016 is different than NHANES 2007-2014, so I change the format for 2015-2016 and using 2007-2014 format.

No issue anymore

(4**) Total nutritional intake:** Someone have dietary day one total energy but not dietary day two, if they have both, then average used, if they have only one, then that value is used.

Yes we do it

(5) **Physical Activity:** NHANES 2011+ does not have: Hours watch TV or videos past 30 days, 0:less than 1 hour 1: 1 hour…..

Remove those variables

(6) **Dietary supplement intake:** Because we have day 1 and day 2 dietary recalls, I combined this information, if any day they have a dietary supplement intake no matter day 1 or day 2, then they are categorized as having dietary supplement intake overall.

Yes

(7) **Multivitamin user:** Because we have day 1 and day 2 dietary recalls, I combined this information, if any day they have a multivitamin intake no matter day 1 or day 2, then they are categorized as having multivitamin intake overall. Query vocabulary: \*VITAMIN\*

Yes

(8) Unit of every variable

(9) Systolic blood pressure has 4 readings, I used average of four readings.

Yes

(10) I think right now the most important question is how to define:

Consumption of fruits

Consumption of Cruciferous vegetables

Consumption of fish

Consumption of unprocessed poultry

Consumption of unprocessed red meat

Consumption of legumes

Consumption of total dairy

Consumption of nuts and

seeds

Consumption of eggs

Consumption of processed meat

Consumption of whole grain

Consumption of whole bread

Consumption of total meat

Consumption of coffee

Consumption of desert/sweet

Consumption of cheese

Consumption of beef

Consumption of methionine

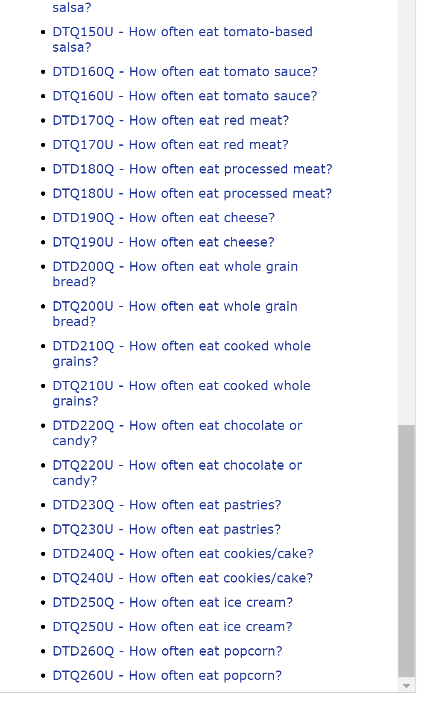
Consumption of berries

Consumption of blood dishes

In terms of short food code description (DRXFCSD):



We also have how often information:



But we will not have gram or kcal information by using food code.

Questions for August 10th meeting:

**Our focus will be age 20-79, this is because some variables only have information for 20+, and age is top coded at 80, which means >= 80 years of age is coded as 80. Also, some people are just ineligible for follow up because of death linkage issues, so there are 21852 participants and 2022 death without considering missing data.**

**First, the problem is that we don’t have food frequency information (except for fish only) because food frequency is only available for NHANES 2009-2010 but not for other years. This means that we can only use Food Patterns Equivalent Database. Let’s discuss food variable first.**

(1) Consumption of fish

In NHANES, we only have frequency information about consumption of fish (that means we don’t have frequency information about consumption of other foods)

In NHANES, we have Shellfish: Clams, crabs, crayfish, lobsters, mussels, oysters, scallops, shrimp, other shellfish, other unknown shellfish eaten

Fish: breaded fish products eaten, tuna, bass, catfish, cod, flatfish, haddock, mackerel, perch, pike, pollock, porgy, salmon, sardines, sea bass, shark, swordfish, trout, walleye, other fish, other unknown fish eaten

On the other hand, in Food Patterns Equivalent Database, we have ounce equivalent information about seafood, which can be combined from seafood high in n-3 fatty acid+ sea food low in n-3 fatty acid.

Which one do we use, in terms of consistency, I suggest only using ounce equivalent and disregard frequency information?

Yes

(2) Total fruit includes citrus, melons, and berries, other fruits and fruit juice

Exclude Fruit juice.

(3) Total vegetables include dark green vegetables, total red and orange vegetables, other red and total starchy vegetables and other vegetables.

No problem here

(4) Consumption of red meat: Can we use Meat variable to define red meat? Meat variable Beef, veal, pork, lamb, and game meat; excludes organ meat and cured meat (oz. eq.) No problem here

(5) We have consumption of legumes

PF\_legumes

(6) Total dairy includes consumption of milk, yogurt and cheese

No problem here

(7) Consumption of processed meat: Can we use cured meat to define processed meat? Cured meat includes Frankfurters, sausages, corned beef, and luncheon meat that

are made from beef, pork, or poultry (oz. eq.)

Yes

(8) We have whole grain

Just use whole grains

(9) We don’t have bread, coffee, desert and sweet, and we remove them

Remove

(10) We have consumption of cheese

Still include them

(11) We have consumption of poultry

No problem

(12) We have consumption of nuts and seeds

No problem

(13) We have consumption of eggs

No problem

(14) We have consumption of total meats: total meats defined as meat+ cured meat+ organ meat + poultry + from seafood high in n-3 fatty acid+ sea food low in n-3 fatty acid.

No problem

(15) We don’t have consumption of beef, methionine, blood dishes in detail

Remove

(16) We don’t have consumption of berries in detail

Remove

(17) Because Meat variable is the summation of beef+pork+lamb+veal+game meat, in order to create a variation in the definition of red meat variable, I created four other meat variable: beef+veal; beef+veal+pork; beef+veal+lamb; beef+veal+pork+lamb. Creating these variables has limitations because we search words like: BEEF, PORK, VEAL, LAMB (case insensitive), however, some food have red meat contained in it but omitted by description of food, such as **Chili con carne with beans**, this food has no BEEF or PORK or VEAL or LAMB in the description but this food is considered having grounded beef and PF\_MEAT is non-zero.

No problem

(18) Can we discuss the missingness:

Age: 20+-79 & eligible for linkage 21852, 2022 death

**ALQ130: 5455 missing. This is the best I can get.**

I combined ALQ120Q - How often drink alcohol over past 12 months and ALQ130 to create a new variable ALQ130.

**BPQ050A: 15000+ missing: taking prescribed medicine for HBP**

I could not do anything about it. I don’t think imputation make sense, so I suggest that we remove this variable.

No problem

**RHQ060(**age at last period missing much than others) premenopausal, postmenopausal

This variable has 50% missing for women, so we cannot create menopausal status effectively. In complete cases, 5000 women will be cut into 2500.

I look into this.

**OCD241: 9326 missing values for occupation group code (to create cleric worker, manual worker…)**, **so I used other variables**

Solution: Type of work done last week have fewest missing values and can use it to combine with Hours worked last week at all jobs to create non-worker, light worker, moderate worker, heavily worker. Then the missing value could be reduced to **1175.** How to create that variable?

Yes

**PAD660: 17195 missing values for Minutes vigorous recreational activities, so I used other variables**

Solution: I used Vigorous recreational activities (Yes or No), and Moderate recreational activities (Yes or No), to create a new variable called Activity (Yes or No), then the missing values could be reduced to **0.**

**Yes**

**SMQ040: 12040 missing so many missing values**

**SMD650: 17024 missing so many missing values**

**Solution:** I will categorize participants into 3 categories, (1) Light or Non-smoker includes people who never smoked regularly, people who smoked less than 100 cigarettes, and people who smoked more than 100 cigarettes but quit smoking (2) Moderate smokers includes people who smoked more than 100 cigarettes and are current smoker and smoked <67 cigarettes a month (3) Heavy smokers includes people who smoked more than 100 cigarettes and are current smoker and smoked>=67 cigarettes a month. In this way, the missing values could be reduced to **50**

* **0 Pack Years – Never Smokers**
* **1 – 20 Pack Years – Light Smokers**
* **1 – 40 Pack Years – Moderate Smokers**
* **More than 40 Pack Years – Heavy Smokers**
* **20 cigarettes per pack**

40 pack years = 67 cigarettes per month

https://www.pdcenterlv.com/blog/what-is-a-cigarette-smoking-pack-year/#:~:text=Pack%20Years%20and%20Smoker%20Classification&text=0%20Pack%20Years%20%E2%80%93%20Never%20Smokers,40%20Pack%20Years%20%E2%80%93%20Heavy%20Smokers

**Complete cases analysis: 1041 death 11006 people**

(19) Refused or Don't Know will be treated as missing values.

(20) Borderline diabetes treated as Yes or No?

(21) 0 Not at all 4163 4163

1 Several days 957 5120

2 More than half the days 234 5354

3 Nearly every day 216 5570 for depression?

(22) 1 Excellent, 654 654

2 Very good, 1747 2401

3 Good, 2631 5032

4 Fair, or 1270 6302

5 Poor? 255 6557

(23) How we quantify OCQ180 and create categories?

(24) Can we calibrate the unit of measure? And can we discuss how to categorize each variable one by one?

(25) How can we conduct specification curve? Any packages or code example for this?

https://cran.r-project.org/web/packages/specr/vignettes/specr.html

The below questions are invalid:

(1) Let’s define dietary variables:

Consumption of fruits

[During the past month], how often did {you/SP} eat fruit? Include fresh, frozen or canned fruit. Do not include juices. [You can tell me per day, per week or per month.]

Q: Do we need to include juices for consumption of fruit?

(2) Consumption of vegetables

Q: How do we define consumption of vegetables? In NHANES, we have DTD090Q - How often eat leafy/lettuce salad?, DTD100Q - How often eat fried potatoes?, How often eat non-fried potatoes?, DTD120Q - How often eat beans?, DTD130Q - How often eat other vegetables? [During the past month], not including what you just told me about [lettuce salads, potatoes, cooked dried beans], how often did {you/SP} eat other vegetables? [You can tell me per day, per week or per month.] It looks like NHANES define vegetables as salad+potatoes+beans+other vegetables?

(3) Consumption of fish

Q: How do we define consumption of fish or we define seafood?

In NHANES, we have Shellfish: Clams, crabs, crayfish, lobsters, mussels, oysters, scallops, shrimp, other shellfish, other unknown shellfish eaten

Fish: breaded fish products eaten, tuna, bass, catfish, cod, flatfish, haddock, mackerel, perch, pike, pollock, porgy, salmon, sardines, sea bass, shark, swordfish, trout, walleye, other fish, other unknown fish eaten

Do we need Fish or We need Fish+Shellfish=Seafood?

We use Yes or No, or we use times/week?

(4) We use times/week, or times/day or times/year? If there are multiple vegetables or fish, are we using the sum? We also have Yes or No information for fish.

(5) Consumption of unprocessed poultry

We do not have this information, unless using individual food and query

(6) Consumption of legume

It looks like legume is approximately beans, and is a type of vegetables, do we still need it? If Yes, then I will use consumption of beans.

(7) Consumption of total dairy

The closest I found is DTD030Q - How often drink milk or on cereal? Can we use it?

(8) Consumption of nuts and seeds

We do not have this information in NHANES unless using individual food and query

(9) Consumption of eggs

We do not have this information in NHANES unless using individual food and query

(10) Consumption of whole grain

I only found consumption of cooked whole grains, including brown rice or other cooked whole grains, such as bulgur, cracked wheat, or millet? But not include white rice. Can we use it?

(11) Consumption of total meat

Looks like we only have red meat + processed meat and seafood (such as bacon, lunch meats, or hot dogs?) Because we do not have chicken, turkey, can we still define total meat?

(12) Consumption of coffee

We have How often drink sweetened coffee/tea? Can we use it?

(13) Consumption of desert/sweet

We have chocolate or candy, pastries, cookies/cake, ice cream, popcorn, Can we combine them to define desert and sweet?

(14) We don’t have Consumption of beef, Consumption of methionine, Consumption of berries, Consumption of blood dishes information. Can we remove them? Only appear once in literature.

Questions for August 12th meeting:

**Continuous or categorical, unit of measure, reference category.**

1. SEQN: Respondent sequence number

* Continuous or categorical: Not applicable
* Unit of measure: NA
* Reference category:NA
* Comment: NA

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DEMO\_F.htm#SEQN

1. RIDAGEYR: Age

* Continuous or categorical: Continuous or categorical
* Unit of measure: years
* Reference category: Refer to other studies. Reference doesn’t matter.
* Comment: As similar as we can as other studies or same interval.

1. RIAGENDER: Sex

* Continuous or categorical: Categorical
* Unit of measure: NA
* Reference category: won’t matter. Then female
* Comment:

1. DMDEDUC2: Educational level

* Continuous or categorical: categorical
* Unit of measure: NA
* Reference category: doesn’t matter
* Comment: use the given categories

1. RIDRETH1: Race

* Continuous or categorical: categorical
* Unit of measure:
* Reference category: doesn’t matter
* Comment: same

1. DMDMARTL: Marital status

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment: same

1. INDFMIN2: Family Income

* Continuous or categorical:
* Unit of measure:
* Reference category: combine there as one categories. Reference doesn’t matter
* Comment: 20+ 20- treated as missing

1. INDFMPIR: Socioeconomic status

* Continuous or categorical:
* Unit of measure:
* Reference category: quntiles
* Comment:

1. SDDSRVIR: Entering cohort year

* Continuous or categorical: categorical
* Unit of measure:
* Reference category:
* Comment:

1. ELIGSTAT: Eligible for death linkage

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

PDF DOCUMENT

1. MORSTAT: All-cause mortality status

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PERMTH\_INT: Follow-up time

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment: **Month? Create another variable using this variable called : age of death or censored.**

1. ALQ130: Alcohol use

* Continuous or categorical: Continuous for now and see other studies how they make categories.
* Unit of measure:
* Reference category:
* Comment: 0 means non-drinkers

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/ALQ\_F.htm

1. BMXBMI: Body mass index

* Continuous or categorical: Continuous for now and see other studies how they make categories.
* Unit of measure:
* Reference category:
* Comment: Implausible BMI? Obesity? Find how other study exclude them

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/BMX\_F.htm#BMXBMI

1. BMXHT: Body height

* Continuous or categorical: Continuous
* Unit of measure:
* Reference category:
* Comment:

1. BMXWT: Body weight

* Continuous or categorical: continuous
* Unit of measure:
* Reference category:
* Comment:

1. BPQ080: History of hypercholesterolemia

* Continuous or categorical: Categorical
* Unit of measure:
* Reference category:
* Comment:

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/BPQ\_F.htm#BPQ080

1. BPQ020: History of hypertension

* Continuous or categorical: categorical
* Unit of measure:
* Reference category:
* Comment:

1. BPXSY: Systolic blood pressure

* Continuous or categorical: Continuous see other studies treat it
* Unit of measure:
* Reference category:
* Comment: quntile

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/BPX\_F.htm

1. DIQ010: History of diabetes

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment: Borderline? See other studies how to treat it, pre-diabetes? If no, then categorize it as NO

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DIQ\_F.htm

1. DPQ020: History of depression

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment: Yes or No? 1,2,3 as YES. 0 as NO

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DPQ\_F.htm#DPQ020

1. HSD010: Health condition

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/HSQ\_F.htm#HSD010

1. CARDIOVASCULAR: History of cardiovascular disease

* Continuous or categorical:
* Unit of measure:
* Reference category:same
* Comment:

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/MCQ\_F.htm#MCQ220

1. MCQ220: History of cancer or malignancy

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. MCQ300C: Family history of diabetes

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. MCQ300A: Family history of myocardial infraction

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. MCQ160F.1: History of stoke

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment: Repeat with CARDIOVASCULAR. Sometimes use stroke, sometimes use cardiovascular disease

1. MENOPAUSAL: Menopausal status

* Continuous or categorical: categorical
* Unit of measure:
* Reference category: same
* Comment:

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/RHQ\_F.htm

1. RHQ540: Hormone therapy users

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. RHQ131: Parity

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. RHQ420: Oral contraceptive use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. RHD143: Are you pregnant now?

* Continuous or categorical:
* Unit of measure:
* Reference category: not put into model
* Comment: Women that are pregnant at baseline, used to exclude pregnant women

1. SLD010H: Sleep

* Continuous or categorical: Continous one and see other studies
* Unit of measure:
* Reference category:
* Comment:

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/SLQ\_F.htm

1. ASPIRIN: Aspirin use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. ATORVASTATIN: Atorvastatin use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. IBUPROFEN: Ibuprofen use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. OPIUM: Opium use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. STATIN: Statin use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. VALSARTAN: Valsartan use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. TKCAL: Total energy intake

* Continuous or categorical: continuous
* Unit of measure:
* Reference category:
* Comment: Extreme value of total energy intake? Refer to other studies.

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DR1TOT\_F.htm#DR1TKCAL

1. TCARB: Total carbohydrates

* Continuous or categorical:
* Unit of measure:
* Reference category: Use as it is and refer to other studies if they calculate % energy. We can do both?
* Comment: % energy? Or mg/day?

1. TFIBE: Total dietary fiber

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment: <https://academic.oup.com/ajcn/article/115/1/189/6329041>
* https://pubmed.ncbi.nlm.nih.gov/9094926/

1. TSFAT: Total saturated fat

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. TMFAT: Total monounsaturated fatty acid

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. TPFAT: Total polyunsaturated fatty acid

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. DRQSDIET: Adoption of special diets

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. TCHOL: Total cholesterol

* Continuous or categorical:
* Unit of measure:
* Reference category: mg not need for conversion to gm
* Comment:

1. TMAGN: Total magnesium

* Continuous or categorical:
* Unit of measure:
* Reference category: mg not need for conversion to gm
* Comment:

1. DSDS: Regular take of nutritional supplement

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DS1TOT\_F.htm#DS1DS

1. MULTIVITAMIN: Multivitamin use

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. OCQ180: Time working

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment: 0 as non-worker. 1-30: part time worker. 30+: full time worker

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/OCQ\_F.htm

1. ACTIVITY: Moderate or vigorous activity

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

1. PAD680: Sedentary lifestyle

* Continuous or categorical:
* Unit of measure:
* Reference category: continuous for now and see other studies
* Comment: Some people include time sleeping? Exclude the one obviously not right, 1000+ not right treat it as missing.

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/PAQ\_F.htm

1. SMOKING: Smoking status

* Continuous or categorical:
* Unit of measure:
* Reference category: same
* Comment:

https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DPQ\_F.htm

1. F\_FRUIT: Total fruits

* Continuous or categorical:
* Unit of measure:
* Reference category: continuous for now and see other studies
* Comment:

PDF DOCUMENT

1. V\_TOTAL: Total vegetables

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PF\_SEAFD: Total seafood

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. G\_WHOLE: Total whole grain

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PF\_MPS\_TOTAL: Total meat

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PF\_MEAT: Total unprocessed red meat

* Continuous or categorical:
* Unit of measure:
* Reference category: treat it as it is and refer to literature
* Comment: Unlikely high daily red meat consumption (>300 g/d)

Won’t for now

1. PF\_CUREDMEAT: Total processed meat

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PF\_POULT: Total poultry

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PF\_EGGS: Total eggs

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PF\_NUTSDS: Total nuts and seeds

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. PF\_LEGUMES: Total legumes

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. D\_TOTAL: Total dairy

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. D\_CHEESE: Total cheese

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. BEEF+VEAL: Total beef + veal

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. BEEF+VEAL+LAMB: Total beef + veal + lamb

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. BEEF+VEAL+PORK: Total beef + veal + pork

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

1. BEEF+VEAL+PORK+LAMB: Total beef + veal + pork + lamb

* Continuous or categorical:
* Unit of measure:
* Reference category:
* Comment:

Questions for August 24th meeting:

We have three things remaining:

**(1) One progress report due on September 30th and**

**(2) Final report on November 14th**

**(3) A Capstone presentation on December 15-16**

**Let’s look at the dataset first. After cleaning, we have 10639 persons and 1018 death.**

**For women, we have 5133, and 385 deaths.**

**For men, we have 5506, and 633 deaths.**

(1) How many specifications are we targeting? In the hurricane paper, it is 1728 and they displayed 300 results. Given that we need to submit a final report in November 14th.

The package will do that. Time is enough.

(1) Are we using time-dependent cox regression or time-independent cox regression, or both? What kind of statistics I need to record? For example, HR, CI, test statistics, P-value, given that we are using specification curve, it looks like we don’t need CI. Or we record everything just in case.

Time-independent cox regression. HR, CI, P-value

(2) Which energy model are we using: (1) The unadjusted model which does not adjust for energy and individual component such as carbohydrates, fats… (both total and remaining). (2) The all-component model that does not adjust for energy, but adjust for individual component (3) The standard model (or so-called residual model or so-called substitution model) that adjust for total energy. (4) The multivariable nutrient density model that calculate nutrient exposure into a percentage of total energy and also adjusted for total energy.

By the way, no one use energy partition model. Given that the estimates (HR) have different interpretations for (1) & (2) (Total causal effect) and (3) (Average relative causal effect) and (4) (Average relative causal effect rescaled as a proportion of total energy), can we display them altogether in specification curve graph? Or we just use (3)

Yes, we will pick two most used model.

(3) How we treat meat variable, continuous or categorical? And food variable? And cholesterol …

For red meat variable, 3 specifications. For food variable, also 3 specifications.

(4) How we interpret results for standard model (substitution model) in terms of ounce equivalent:

For example, y=a\*red meat + b\*total energy

For example, y=a\*red meat + b\*fruit + c\*total energy

This model could also control for total meat intake instead of total energy? Are we doing total meat intake?

We don’t use ounce, we covert in grams. Only consider total energy for standard model.

(5) The standard model (substitution model) and density model could both do leave one out analysis? Do we do leave one out?

We don’t do leave one out.

(6) Which interaction are we going to examine?

Age, sex and BMI.

No interaction.

Interaction with age

Interaction with Sex

Interaction with BMI

(7) Which subgroup analysis are we going to examine?

Age Group

No subgroups for age

20-30

30-40

…

…

Sex

No subgroups sex

Male

Female

(8) Do we need to examine proportional hazard assumption each time we run a specification?

No.

(9) Discuss about what is linear trend? Restricted cubic splines?

We don’t do these.

(10) Do we do regression calibration?

We also don’t do that.

(11) Do we do imputation or replacing missing value with median? We already discuss that we do complete cases analysis.

We don’t do.

(12) Definition of total energy

I define total energy = Carbohydrate (gm) + Dietary fiber (gm) + Total saturated fatty acids (gm) + Total monounsaturated fatty acids (gm) + Total polyunsaturated fatty acids (gm) + Cholesterol (mg) + Total meat (Red meat + Poultry + Seafood) (ounce equivalent)

We are missing consumption of sugar (not included in any paper); do we need to bring gm or mg to kcal.

We need to look at the dataset and discuss. Do we need to bring ounce equivalent or cup equivalent to kcal? Gram, cup we leave as it is.

Which dietary variables could not be included at the same time? For example, we can include Carbohydrate (gm) + Dietary fiber (gm) + Total saturated fatty acids (gm) + Total monounsaturated fatty acids (gm) + Total polyunsaturated fatty acids (gm) + Cholesterol (mg) + Total meat (Red meat + Poultry + Seafood) (ounce equivalent). Or we can include red meat + poultry + seafood + eggs + nuts and seeds + total diary ….

We can include both.

(13) For other variables, how we decide which one to include or not. For example, BMI, Annual family income and smoking. We could have a standard list of non-dietary variables that we always include, and an additional list of groups of variables we add into standard lists once at a time.

Age, sex, smoking. Chirag vibration of effect paper. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4555355/>

Red meat

Contious

Qunitles

Qurtiles

Variables

Age + SEX+ SMOKING+ Total energy+ Red meat

Exercise

Physical activity

….

….

…

….

(14) We also have variation in definition of red meat variable. BEEF+VEAL

BEEF+VEAL+PORK

BEEF+VEAL+LAMB

BEEF+VEAL+PORK+LAMB

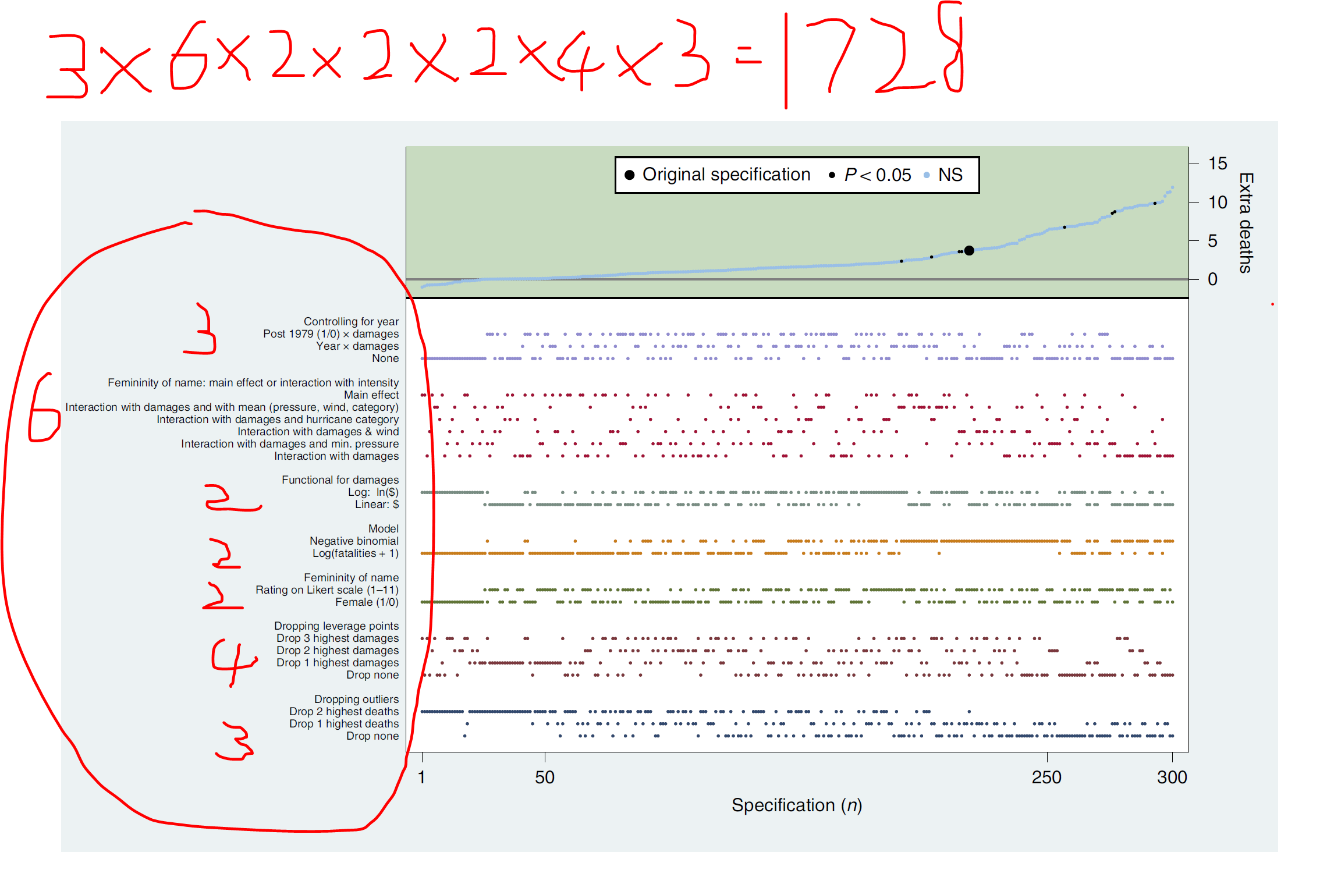
BEEF+VEAL+PORK+LAMB+GAME MEAT

I suggest we don’t consider it because the calculation is wrong.

Just disregard it

**Questions about specification curve:**

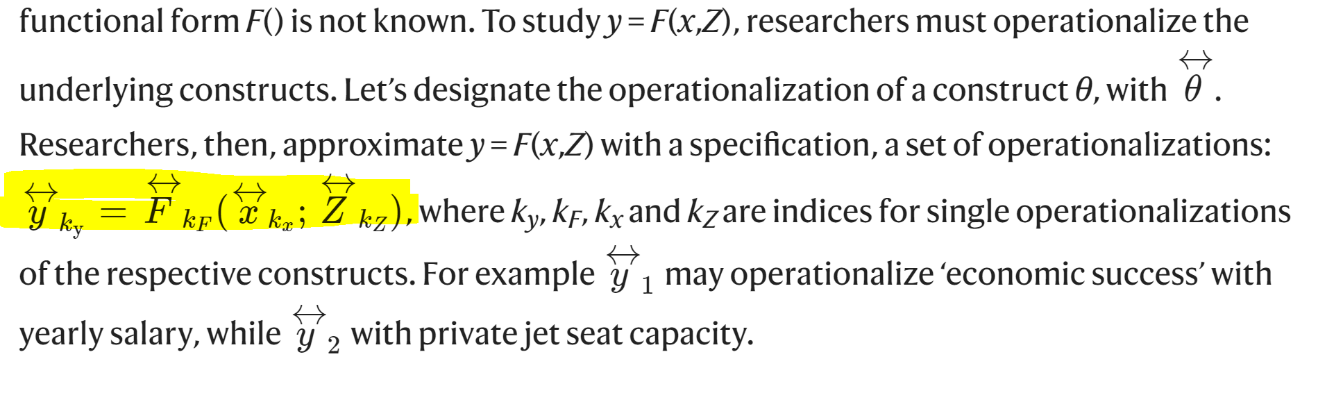
(1) We need to identify the set of theoretically justified, statistically valid and non-redundant specifications, in other words the left panel of the specification graph. Refer to analytical decisions document and we need to discuss. Can we be sure they are non-redundant with other specifications?



We need a complete list of specifications. We need to draw a reasonable template of left panel.

(2) What is consequential? Using the above graph. My understanding is that the analytical decisions or trends that severely affect the effect estimates and signs?

(3) Let’s discuss operationalizations of the respective constructs in our reseach:

****

(4) It looks like we have three statistics to conduct joint statistical inference (considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?): median effect size, proportion of significant results, and average Z value, which one do we use? The author propose report both the second and third.

Median effect size. Work around, package.

(5) There are two separate categories for joint inference: experiment data joint inference and non-experimental data joint inference. Is our data experimental data? We are observational data by the way. I suspect we could not directly shuffle red meat consumption variable 500 times to get under the null of no effect. Because for non-experimental data, covariates are possibly correlated with the predictor of interest (red meat).

(6) The conduct of non-experiment data joint inference looks like not applicable to cox regression. How to Generate K different dependent variables under the null in cox regression? What’s dependent variable in cox regression by the way?

(7) Need the code for cox regression and the code for plotting specification graph.

https://philippmasur.de/2020/01/02/how-to-do-specification-curve-analyses-in-r-introducing-specr/

**Template:**

Energy models (2)

* The Standard (or so-called Residual) model (Relative causal effect)
* The multivariable nutrition density model (Relative causal effect rescaled as a proportion of total energy)

Adjusting variables (249=5.629\*1014)

* Standard lists: red meat, age (continuous), sex, smoking, total energy, if only women selected, then menopausal status, use of hormone therapy, parity, oral contraceptive use is automatically adjusted.
* + Years of entering cohort
* + Race/Ethnicity
* + Education
* + Martial Status
* + Alcohol consumption (continuous)
* + Alcohol consumption (categorical)
* + Occupation
* + Physical activity
* + Sedentary lifestyle
* + Sleep hours
* + Annual family income
* + Socioeconomic status
* + BMI (continuous)
* + BMI (categorical)
* + Systolic blood pressure
* + General health condition
* + History of hypercholesterolemia
* + History of hypertension
* + History of diabetes
* + History of depression
* + History of cardiovascular disease
* + History of cancer
* + Family history of diabetes
* + Family history of myocardial infraction
* + Use of Aspirin
* + Use of Ibuprofen
* + Use of Opium
* + Use of statin
* + Use of Valsartan
* + On special diet
* + Dietary supplement use
* + Cured meat
* + Poultry
* + Fruit
* + Total Vegetables
* + Seafood
* + Whole grains
* + Eggs
* + Nuts and Seeds
* + Legumes
* + Total Diary
* + Total Carbohydrates
* + Total Dietary Fiber
* + Total saturated fatty acids
* + Total monounsaturated fatty acids
* + Total polyunsaturated fatty acids
* + Total Cholesterol
* + Total Magnesium

Variable Type of red meat (3)

* Quintiles
* Quartiles
* Continuous

Interactions (4)

* No interaction included
* Include interaction only with age
* Include interaction only with sex
* Include interaction only with BMI

Subgroups (Age) (7)

* All age
* 20-29 years old
* 30-39 years old
* 40-49 years old
* 50-59 years old
* 60-69 years old
* 70-79 years old

Subgroups (Sex) (3)

* All sex
* Male
* Female

Questions for September 2th meeting:

We now have 2\*5.629\*3\*4\*7\*3\*10^14=645\*10^14 specifications, is it too much?

(1) Can we decide if we want to use energy density model? If yes, then which variables are we going to divide by total energy? Just red meat. 100/2000 Gram/kcal increase continuous

100 Gram increase if continuous

Highest vs. lowest reference category.

(2) Can we decide the template and reduce specifications? Do we need to combine something? Or delete something? We do random sampling of all variable combinations

[**https://gs.mcmaster.ca/program/health-research-methodology/**](https://gs.mcmaster.ca/program/health-research-methodology/)

Questions for September 9th meeting:

We now could make 1680 specifications (1 meat variable type \* 2 models \* 4 interactions \* 10 combinations of adjusting variables \* 3 sex \* 7 age) so far, treating meat variable as continuous and used two models: standard model and density model.

(1) Convergence problem sometimes does not converge, leading to HR to be very very big (10^6) and sometimes converge but the confidence interval is huge. By the way, do we need to plot confidence interval in the graph, I suspect not, because we are using different models (standard, density) and the CI is not comparable in graph.

sparse data bias

https://www.bmj.com/content/352/bmj.i1981#:~:text=Sparse%20data%20bias%3A%20the%20bias,total%20sample%20size%20appears%20large.

(2) Technical question: Is this how we do cox regression model

Coxph(Surv(follow-up,event)~variables, data=Data, ties=”Breslow”)?

I will use default

(3) Technical question:

How can I include the meat variable as quartile variable in cox. It looks like I could not factorize the variable and include it directly in run\_spec() function because this function could not extract 4th quartile coefficient vs reference automatically. In the case, my understanding is to create 3 binary variables, X1, X2, X3 each denote the person is in 4th, 3rd, 2nd quartile and not and include it in the model so that the run\_spec() could directly extract X1 coefficient and that what we want.

What I achieved so far:

* I examined meat variable as continuous and two models: standard model and density model. Looks like the standard error is different in these two models?
* Included interaction, here one thing needs to pay attention to is when we are doing subgroup analysis for sex, for example Female, including interaction with sex is meaningless.
* I created a function that could randomly sample variables and include them in the model as adjusting variables. Number in the graph correspond to variable names and I have a list of all variables labeled as 1-48. I also added a complexity that alcohol (continuous) and alcohol (categorical) will not be included at the same time and BMI (continuous) and BMI (categorical) will not be included at the same time.
* For female subgroup analysis, we also adjust for menopausal status, oral contraceptive use, parity and hormone therapy use. For male and all sex, these are not adjusted for cox regression.

Next thing to do:

* Treat meat variable as quartile and quintiles
* Rerun two models

Questions for 9.16 meeting:

**5000 specifications need 10 minutes for my computer to run.**

(1) How to interpret HR when there is an interaction term?

For example: Denote X to be quartile of red meat

ln(Hazard)=b0X2+b1X3+b2X4+b3X2\*BMI+b4X3\*BMI+b5X4\*BMI+b6\*BMI+….

Then the hazard of the 4th quartile will be

ln(Hazard)=b2+b5\*BMI+b6\*BMI+……

The reference quartile will be

ln(Hazard)=b6\*BMI

Then the hazard ratio will be ln(HR)=b2+b5\*BMI where b2 is the coefficient corresponds to 4th quartile and b5 is the coefficient corresponds to interaction term of 4th quartile and BMI. It looks like the **HR=exp(b2+b5\*BMI)** changes with BMI when comparing 4th quartile with lowest quartile when there is an interaction term, in this case, can we still plot HR (actually we plotted exp(b2)… disregarded b5\*BMI)? Do we still need to include interaction term for sex and age?

In other words, if there is a interaction term, that means HR changes over age, sex, BMI so there is no single value to be plotted in specification graph.

Remove interaction.

(2) What are the green lines? This is because all the lines are stacked. Need to make graph bigger then the lines could be separated and green line disappear.

This not a problem.

(3) Ties: a character string specifying the method for tie handling. If there are no tied death times all the methods are equivalent. Nearly all Cox regression programs use the Breslow method by default, but not this one. The Efron approximation is used as the default here, it is more accurate when dealing with tied death times, and is as efficient computationally. The “exact partial likelihood” is equivalent to a conditional logistic model, and is appropriate when the times are a small set of discrete values. See further below.

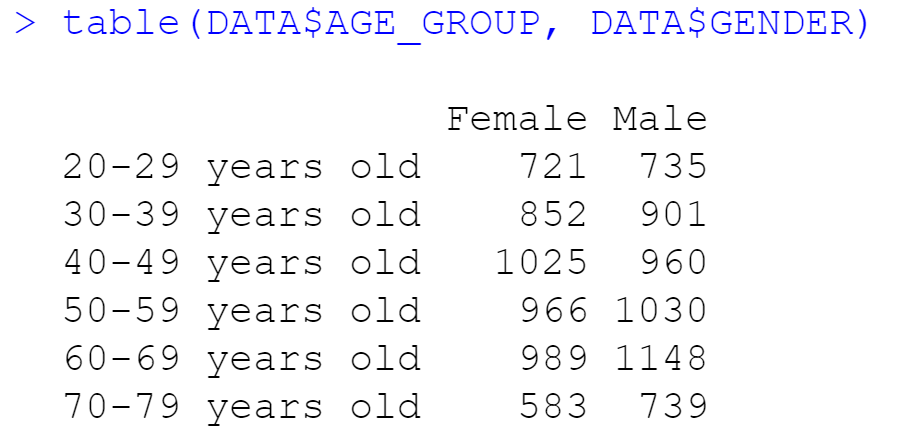
We will use the efron. Default.

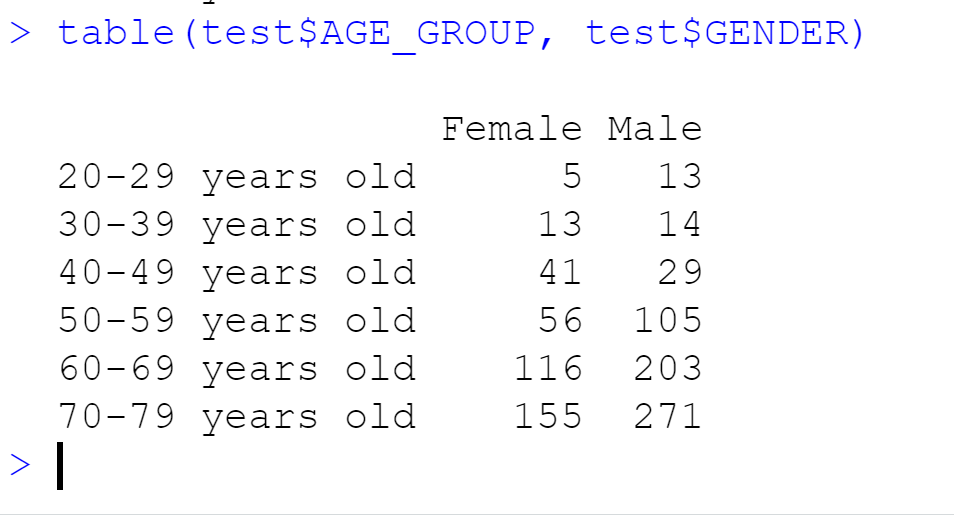
(4) Let’s check the term age in cox model, I used continuous version, is it ok? Or do we need age\_group?

Use continuous age

(5) The distribution of death is below: We have 10649 people and 1021 death.

The distribution of total people:





Combine into 20-39, 40-59, 60-79

(6) It looks like HR could be very large (convergence problem) and even if it is small, confidence interval would be very large which makes the graph unreadable if we also plot confidence interval. For example, HR=0.4 which is reasonable but CI=(0.05,33) , 33 is so large that if we plot that on the graph, then all other HR would centered (flat line) around 1 and the graph is ugly.

(7) What is a good HR range 0.2-2, What is a reasonable CI range 0-3.5?

CI:0.2-5 use CI to filter all results

(8) **Error messages 1:**

1: Problem while computing `res = map2(.data$model, formula, ~do.call(.x, list(data = df, formula =.y)))`.

i Loglik converged before variable 5,6,7,8,24 ; coefficient may be infinite.

41: Problem while computing `res = map2(.data$model, formula, ~do.call(.x, list(data = df, formula =.y)))`.

i Ran out of iterations and did not converge

**Error message 2:**

not finite in coxph.wtest

UNPROCESSED\_RED\_MEAT\_DENSITY\_QUINTILE\_5th

-271.1683648

EDUCATION9-11th Grade (Includes 12th grade with no diploma)

-106.3790473

EDUCATIONHigh School Grad/GED or Equivalent

-202.4553133

EDUCATIONSome College or AA degree

-109.5290369

EDUCATIONCollege Graduate or above

-86.4583873

MARTIAL\_STATUSMarried

515.6514415

MARTIAL\_STATUSWidowed

534.0874191

MARTIAL\_STATUSDivorced

495.9344904

MARTIAL\_STATUSSeparated

249.2214402

MARTIAL\_STATUSLiving with partner

522.8473790

ALCOHOL\_CONTINOUS

35.8746496

OCCUPATIONPart time worker

-122.8047304

OCCUPATIONFull time worker

25.5523000

SLEEP>=9 hours/night

207.0075789

SLEEP5-8 hours/night

64.3375233

FAMILY\_INCOME$15,000 to $34,999

-283.4929370

FAMILY\_INCOME$35,000 to $64,999

-307.2908428

FAMILY\_INCOME$65,000 to $100,000 and Over

-91.8447325

BMI\_CONTINOUS

6.8562338

SYSTOLIC\_BLOOD\_PRESSURELower-middle

-174.6085217

SYSTOLIC\_BLOOD\_PRESSUREMiddle

-128.3704400

SYSTOLIC\_BLOOD\_PRESSUREUpper-middle

12.4184518

SYSTOLIC\_BLOOD\_PRESSUREHigh

-425.5585885

HISTORY\_OF\_HYPERCHOLESTEROLEMIAYes

-14.0905636

HISTORY\_OF\_HYPERTENSIONYes

99.8302771

HISTORY\_OF\_DIABETESYes

82.6452140

HISTORY\_OF\_CARDIOVASCULAR\_DISEASEYes

393.6891371

FAMILY\_HISTORY\_OF\_MYOCARDIAL\_INFRACTIONYes

210.1669080

IBUPROFENYes

-370.2862655

OPIUMYes

544.2334227

ON\_SPECIAL\_DIETYes

-180.2163435

DIETARY\_SUPPLEMENTYes

-64.1900147

PROCESSED\_MEAT

-0.4761753

POULTRY

0.1104745

FRUITS

-51.0801772

VEGETABLES

-68.7595193

SEAFOOD

-0.6195621

EGGS

0.3482046

TOTAL\_DAIRY

56.0156608

CARBOHYDRATES

0.8957703

MONOUNSATURATED\_FATTY\_ACID

2.3392483

UNPROCESSED\_RED\_MEAT\_DENSITY\_QUINTILE\_2nd

-76.4107756

UNPROCESSED\_RED\_MEAT\_DENSITY\_QUINTILE\_3rd

-139.8655151

UNPROCESSED\_RED\_MEAT\_DENSITY\_QUINTILE\_4th

-54.1545541

AGE\_CONTINIOUS

-18.3156700

GENDERMale

0.0000000

SMOKINGModerate smoker

267.1594203

SMOKINGHeavy smoker

296.6299216

TOTAL\_ENERGY

-0.0736314

MENOPAUSAL\_STATUSPostmenopausal

-338.1229130

HORMONE\_THERAPY\_USEYes

59.4189208

PARITYParous

-97.1931570

ORAL\_CONTRACEPTIVE\_USEYes

132.1679087

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15]

[1,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[2,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[3,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[4,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[5,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[6,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[7,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[8,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[9,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[10,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[11,] NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN 0.00 0.00000 0.000000 0.000000

[12,] 0 0 0 0 0 0 0 0 0 0 0 38609.79 0.00000 0.000000 0.000000

[13,] 0 0 0 0 0 0 0 0 0 0 0 0.00 1.39719 0.000000 0.000000

[14,] 0 0 0 0 0 0 0 0 0 0 0 0.00 0.00000 1.444429 0.000000

[15,] 0 0 0 0 0 0 0 0 0 0 0 0.00 0.00000 0.000000 1.438164

[16,] 0 0 0 0 0 0 0 0 0 0 0 0.00 0.00000 0.000000 0.000000

[17,] 0 0 0 0 0 0 0 0 0 0 0 0.00 0.00000 0.000000 0.000000

[18,] 0 0 0 0 0 0 0 0 0 0 0 0.00 0.00000 0.000000 0.000000

[,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]

[1,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[2,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[3,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[4,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[5,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[6,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[7,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[8,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[9,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[10,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[11,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[12,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[13,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[14,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[15,] 0.000000 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[16,] 1.429136 0.000000 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[17,] 0.000000 1.476943 0.000000 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[18,] 0.000000 0.000000 1.396534 0.000000000 0.00000 0.00000 0.000000 0 0.000000 0.00000

[,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34]

[1,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[2,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[3,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[4,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[5,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[6,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[7,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[8,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[9,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[10,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[11,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[12,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[13,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[14,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[15,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[16,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[17,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[18,] 0.0000 0.000 0.000000 0 0 0.0000 0.00000 0.00000000 0.000000e+00

[,35] [,36] [,37] [,38] [,39] [,40] [,41] [,42] [,43]

[1,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[2,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[3,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[4,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[5,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[6,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[7,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[8,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[9,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[10,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[11,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[12,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[13,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[14,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[15,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[16,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[17,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[18,] 0.00000 0.000000 0.000000000 0.000000000 0.0000000 0.000000e+00 0.000000000 0.000000 0.00000

[,44] [,45] [,46] [,47] [,48] [,49] [,50] [,51] [,52] [,53]

[1,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[2,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[3,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[4,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[5,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[6,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[7,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[8,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[9,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[10,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[11,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[12,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[13,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[14,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[15,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[16,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[17,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[18,] 0.00000 0.0000000 0 0.0000 0.00000 0.000000e+00 0.0 0.0000 0.000000 0.000000

[ reached getOption("max.print") -- omitted 35 rows ]

(9) I will check the requirement of writing.

(1) Table of characteristics.

(2) flow chart.

(3) Share point.

Questions for 9/26 meeting:

(1) After we group people into smaller age group: 20-39,40-59,60-79. The error warning messages 1 reduced from 144 to 16 and error messages 2 disappear.

(2) It takes 5 minutes to run 1440=2\*3\*20\*3\*4 specifications, 222 specifications are dropped because CI lower bound<0.2 or CI upper bound>5. The drop rate is 15.4% and that means 85% of the results are kept in the graph.

(3**) In September 30, we need to submit a progress report.** The final progress report must indicate which journal the student and mentor have selected for the project and a link to the “Instructions for Authors” must be included.

**Need a title and targeted journal. Let’s look at progress report 2.**

(4) **Students submit a capstone report in the form of a full‐length journal article manuscript for their targeted journal.** The Capstone Oversight Committee may request revisions.

Submission of Final Report

‐ Draft a final report and receive approval from your mentor.

- Submit report as .pdf or .doc to Canvas and Turnitin

‐ Receive feedback from Capstone Oversight Committee that needs to be incorporated into the final report.

Deadlines - 48-credit MBI, Class of 2023: **November 14th**

Report Format

**The completed capstone paper should be consistent with the requirements of the specific journal to which it could be submitted and will generally include the following sections:**

**● Introduction**

**● Methods**

**● Results**

**● Discussion**

**● References**

**● Acknowledgments**

The mentor must approve the completed written manuscript in advance of submission. Mentors may be contacted at any time to verify that they have reviewed and approved a submitted report. It is strongly encouraged that students submit their manuscript to the selected journal**. Students also have the option to submit their final report to DASH –Digital Access to Scholarship at Harvard – which is a central, open-access repository of research by members of the Harvard community. Students must have permission of their project mentor to deposit to DASH. Learn more at dash.harvard.edu.**

(5) Can you give me some examples of

* + Table of characteristics.
  + flow chart.
  + Share point.

**Meeting on 10/14:**

Q1: What happens if sum of percentage=101 or 99% in characteristic table after rounding? Such as education.

Fine.

Q2: For sedentary lifestyle (hours) and socioeconomic status (PIR) and systolic blood pressure (mmhg), we manually divide them into fifth categories, so that each group contain same amount and percentage, should we report that or should we report the continuous version?

Use continuous version.

Q3: We used month of follow-up, do we need to convert to year of follow up, and subsequently refit specification curve?

Don’t need to covert.

Q4: For continuous variables, I used median (q10,q90) because data are right skewed, I did not use mean (sd), for categorical variables, I used n (%).

Q5: Which block in the manuscript I am responsible for? What suggestions in writing those blocks and can you provide example paragraph that I can mimic?

Finish method before next meeting.

Q6: We are required to put the manuscript in a target journal format? Have we decided which journal? Is time enough to put everything in a journal format given the due date of 11/14, do we need to decide now?

Clinical

Q7: suggestions for flow chart?

Meeting on Oct.28:

(1) Schoenfeld residuals for 5 models: two negative significant, two positive significant, one non-significant

(2) Analytical decisions for 20 primary studies

(3) Do not need to do the results for 20 primary studies, dena will do it.

(4) Fill in the XX in main results