Understanding Strengths & Limitations of Secondary Data in Gender/Sex Epidemiological Analysis

Developed in Women, Gender and Health 207:
Advanced Topics of Women, Gender, and Health,
Harvard School of Public Health, Spring 2024

Teaching Example Authored by
Emily Newton-Hoe, Meekang Sung, & Seetha Davis

Week 8 Wednesday

Using gender/sex variables in secondary data analysis

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

Teaching example made in WGH 207,
Emily Newton-Hoe, Meekang Sung, & Seetha Davis

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

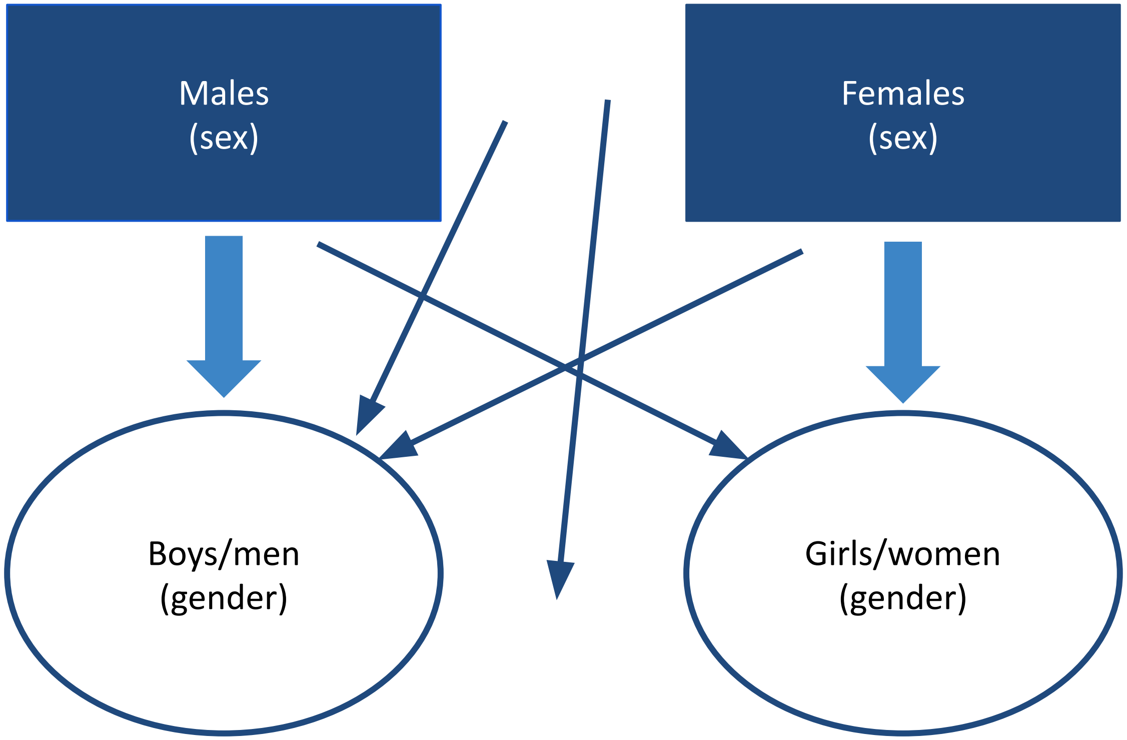
Sex? Gender?

- **Sex:** biological construct premised upon biological characteristics enabling sexual reproduction.
    - characteristics: secondary sex-characteristics, gonads, or sex chromosomes
    - categories: male, female, intersex

- **Gender:** social construct regarding culture-bound conventions, roles, and behaviors for, as well as relations between and among, women and men and boys and girls.
    - both gender relations and **biologic expressions of gender** vary within and across societies
    - often thought of in terms of gender identity and gender expression

Krieger N. Genders, sexes, and health: what are the connections—and why does it matter? International Journal of Epidemiology. 2003 Aug;32(4):652–7.

Sex? Gender?



Males
(sex)

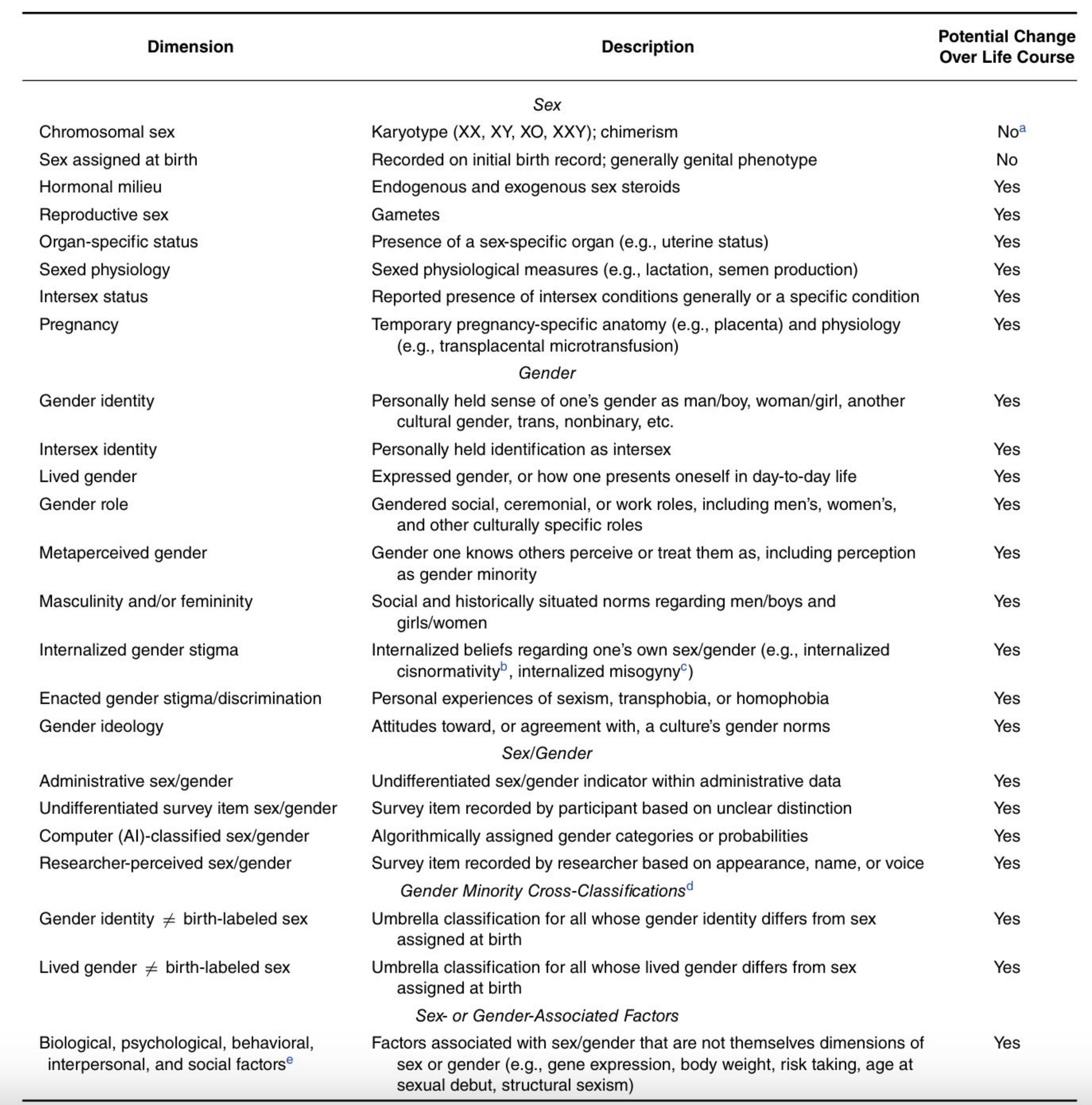
Females
(sex)

Boys/men
(gender)

Girls/women
(gender)

Katz-Wise S, Todd K. WGH 207 2024 S1 Slides

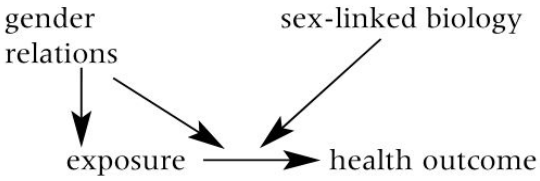
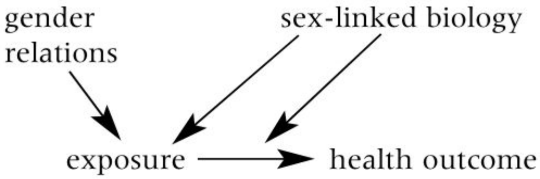
Multidimensionality of sex and gender



Dimension	Description	Potential Change Over Life Course
	Sex	
Chromosomal sex	Karyotype (XX, XY, XO, XXY); chimerism	No[a]
Sex assigned at birth	Recorded on initial birth record; generally genital phenotype	No
Hormonal milieu	Endogenous and exogenous sex steroids	Yes
Reproductive sex	Gametes	Yes
Organ-specific status	Presence of a sex-specific organ (e.g., uterine status)	Yes
Sexed physiology	Sexed physiological measures (e.g., lactation, semen production)	Yes
Intersex status	Reported presence of intersex conditions generally or a specific condition	Yes
Pregnancy	Temporary pregnancy-specific anatomy (e.g., placenta) and physiology (e.g., transplacental microtransfusion)	Yes
	Gender	
Gender identity	Personally held sense of one's gender as man/boy, woman/girl, another cultural gender, trans, nonbinary, etc.	Yes
Intersex identity	Personally held identification as intersex	Yes
Lived gender	Expressed gender, or how one presents oneself in day-to-day life	Yes
Gender role	Gendered social, ceremonial, or work roles, including men's, women's, and other culturally specific roles	Yes
Metaperceived gender	Gender one knows others perceive or treat them as, including perception as gender minority	Yes
Masculinity and/or femininity	Social and historically situated norms regarding men/boys and girls/women	Yes
Internalized gender stigma	Internalized beliefs regarding one's own sex/gender (e.g., internalized cisnormativity[b], internalized misogyny[c])	Yes
Enacted gender stigma/discrimination	Personal experiences of sexism, transphobia, or homophobia	Yes
Gender ideology	Attitudes toward, or agreement with, a culture's gender norms	Yes
	Sex/Gender	
Administrative sex/gender	Undifferentiated sex/gender indicator within administrative data	Yes
Undifferentiated survey item sex/gender	Survey item recorded by participant based on unclear distinction	Yes
Computer (AI)-classified sex/gender	Algorithmically assigned gender categories or probabilities	Yes
Researcher-perceived sex/gender	Survey item recorded by researcher based on appearance, name, or voice	Yes
	Gender Minority Cross-Classifications[d]	
Gender identity ≠ birth-labeled sex	Umbrella classification for all whose gender identity differs from sex assigned at birth	Yes
Lived gender ≠ birth-labeled sex	Umbrella classification for all whose lived gender differs from sex assigned at birth	Yes
	Sex- or Gender-Associated Factors	
Biological, psychological, behavioral, interpersonal, and social factors[e]	Factors associated with sex/gender that are not themselves dimensions of sex or gender (e.g., gene expression, body weight, risk taking, age at sexual debut, structural sexism)	Yes

Bauer GR. Sex and Gender Multidimensionality in Epidemiologic Research. American Journal of Epidemiology. 2023 Jan 6;192(1):122–32.

Multidimensionality of sex and gender on exposure-outcome association

		Relevance of:	
Diagrammed illustration	**Exposure—outcome association**	Gender relations	Sex-linked biology
 gender relations sex-linked biology exposure health outcome	Geographical variation in women's rates of unintended pregnancy as linked to variation in state policies re family planning[19]	Yes: for exposure and once exposed	Yes: once exposed
 gender relations sex-linked biology exposure health outcome	Earlier age of human immunodeficiency virus infection among women compared with heterosexual men (in the US)[20]	Yes: for exposure	Yes: for exposure and once exposed

Krieger N. Genders, sexes, and health: what are the connections—and why does it matter? International Journal of Epidemiology. 2003 Aug;32(4):652–7.

PollEv question

After adjusting for known confounding, investigators find that there is no association between gender and depression. Therefore, there is no need to evaluate whether there is effect measure modification by dimensions of gender identity, expression, or other gendered phenomena.

- True
- False

HAVE A GOOD WEEK

EPI 202 Lab 2 Practice Problems

PART I. Data analysis problems using statistical software

In this section, you will analyze data from the National Health and Nutrition Examination Survey (NHANES). This survey examines the health and nutritional status of children and adults in the United States. It has been continuously administered from 1999 to present, with questions capturing demographic information, dietary habits, and other health-related behaviors. You can read more about NHANES on the following website: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

In this exploratory analysis, you will evaluate the relationship between respondent-reported binary "gender" and depression, using data from the 2011-12 NHANES wave. As of the 2023 wave, NHANES continues to collect data on gender as a binary variable, which has been consistent since the first wave in 1999.

The relevant variables in the dataset are described below:

Variable Name	Description
riagendr	"Gender of the participant" – binary options only. 1 = male, 2 = female.
dpq_score	Total depression score, 0-27, with higher scores indicating more severe depression.
ocd150	Type of work done last week 1 = working at a job or business 2 = with a job or business but not at work 3 = looking for work 4 = not working at a job or business

The dataset name is nhanes_1112.csv and is available for download from Canvas. Please use R for this analysis. Sample R code for the calculations is included for each question.

1. Calculate the prevalence ratio for the association between "gender" and prevalence of depression. Interpret this ratio.

```
setwd("INSERT FILE PATH HERE")
nhanes <- read.csv("nhanes_1112.csv")

# Recode continuous depression into binary depression score, using score >= 10 as
cutoff
nhanes$depr_2cat <- ifelse(nhanes$DPQ_SCORE>=10, 1, 0)
nhanes$depr_2cat <- factor(nhanes$depr_2cat, levels=c(0,1), labels= c("none or mild
depression", "moderate or severe depression"))

## Question 1: Prevalence ratios of depression by "gender"
# create 2x2 table with "gender" x depression
nhanes$RIAGENDR <- factor(nhanes$RIAGENDR, levels = c(1,2), labels = c("male",
"female"))
```

```
table(nhanes$RIAGENDR, nhanes$depr_2cat)

# prevalence by gender, using table values
prev_female <- 284 / (284+2162)
prev_female

prev_male <- 165 / (165+2324)
prev_male

# prevalence ratio
prev_female / prev_male
```

2. You realize this crude analysis may not fully capture the relationship between respondent gender and depression. Draw a DAG that may more accurately capture the effect by considering the role of gender relations and/or sex-linked biology, as relevant. You may consider looking at Table 2 in this paper for inspiration, and the NHANES codebook to see what other variables are available to you.

3. You decide to re-run your analysis, now including occupation as a potential stratifying variable for the association between respondent gender and depression. Calculate the prevalence of reporting depression by gender separately by level of occupation. Comment on what you observe, including how this analysis leads to different conclusions compared to your answer in question 1.

```
### Question 3: re-run analysis, using occupational status

# recode occupation into working or looking for work vs. not working
nhanes$occ_2cat <- ifelse(nhanes$OCD150<=3, 1, 0)
nhanes$occ_2cat <- factor(nhanes$occ_2cat, levels=c(0,1), labels=c("no paid work",
"paid work or looking for paid work"))

# look at depression by occupation
table(nhanes$occ_2cat, nhanes$depr_2cat)

# create subsets of the data for each "gender"
female_only <- subset(nhanes, RIAGENDR=="female", select=c(1:7))
table(female_only$occ_2cat, female_only$depr_2cat)

male_only <- subset(nhanes, RIAGENDR=="male", select=c(1:7))
table(male_only$occ_2cat, male_only$depr_2cat)

# create subsets of the data for each occupational status
work_only <- subset(nhanes, occ_2cat=="paid work or looking for paid work",
select=c(1:7))
table(work_only$RIAGENDR, work_only$depr_2cat)

nowork_only <- subset(nhanes, occ_2cat=="no paid work", select=c(1:7))
table(nowork_only$RIAGENDR, nowork_only$depr_2cat)
```

For future reading:

Christiansen, D. M., McCarthy, M. M., & Seeman, M. V. (2022). Understanding the influences of sex and gender differences in mental disorders. *Frontiers in Psychiatry*, *13*, 984195. doi: 10.3389/fpsyt.2022.984195

For more information about courses that emphasize gender/sex analysis:

https://www.hsph.harvard.edu/women-gender-and-health/info-sheet/

EPI 202 Lab 2 Practice Problems – Teaching Fellow Guide

PART I. Data analysis problems using statistical software

In this section, you will analyze data from the National Health and Nutrition Examination Survey (NHANES). This survey examines the health and nutritional status of children and adults in the United States. It has been continuously administered from 1999 to present, with questions capturing demographic information, dietary habits, and other health-related behaviors. You can read more about NHANES on the following website: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

In this exploratory analysis, you will evaluate the relationship between respondent-reported binary "gender" and depression, using data from the 2011-12 NHANES wave. As of the 2023 wave, NHANES continues to collect data on gender as a binary variable, which has been consistent since the first wave in 1999.

The relevant variables in the dataset are described below:

Variable Name	Description
riagendr	"Gender of the participant" – binary options only. 1 = male, 2 = female.
dpq_score	Total depression score, 0-27, with higher scores indicating more severe depression.
ocd150	Type of work done last week 1 = working at a job or business 2 = with a job or business but not at work 3 = looking for work 4 = not working at a job or business

The dataset name is nhanes_1112.csv and is available for download from Canvas. Please use R for this analysis. Sample R code for the calculations is included for each question.

1. Calculate the prevalence ratio for the association between "gender" and prevalence of depression. Interpret this ratio.

```
setwd("INSERT FILE PATH HERE")
nhanes <- read.csv("nhanes_1112.csv")

# Recode continuous depression into binary depression score, using score >= 10 as
cutoff
nhanes$depr_2cat <- ifelse(nhanes$DPQ_SCORE>=10, 1, 0)
nhanes$depr_2cat <- factor(nhanes$depr_2cat, levels=c(0,1), labels= c("none or mild
depression", "moderate or severe depression"))

### Question 1: Prevalence ratios of depression by "gender"
# create 2x2 table with "gender" x depression
nhanes$RIAGENDR <- factor(nhanes$RIAGENDR, levels = c(1,2), labels = c("male",
"female"))
```

```
table(nhanes$RIAGENDR, nhanes$depr_2cat)

# prevalence by gender, using table values
prev_female <- 284 / (284+2162)
prev_female

prev_male <- 165 / (165+2324)
prev_male

# prevalence ratio
prev_female / prev_male
```

Interpretation: Survey respondents who self-identified as female had 1.75 times the prevalence of moderate to severe depression compared to survey respondents who self-identified as male.

Relevant 2x2 table:

	None/mild depression	Mod/severe depression	Total
Male	2324	165	2489
Female	2162	284	2446
Total	4486	449	4935

Prevalence among males: 165 / (165 + 2324) = 0.066 → 6.6%
Prevalence among females: 284 / (284+2162) = 0.116 → 11.6%
Prevalence ratio = 0.116 / 0.066 = 1.751

Additional TF notes:

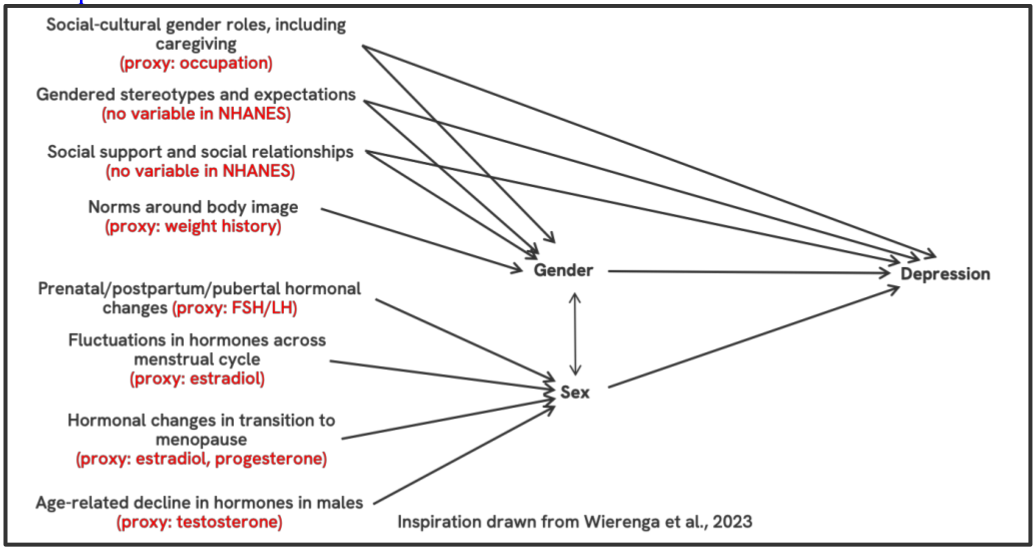
- Emphasize the **importance of being clear when using language** referring to biologic sex (male/female) versus gender (man/woman). Here, the NHANES survey developers seem to be using gender to refer to sex, and thus are using sex-related labels. This is not best practice.
- Highlight that interpretations should be **as specific as possible regarding how sex and/or gender were defined**. Here, the survey was not specific in defining gender, so it is unclear whether people self-identified according to gender identity or gender expression (gender) or if instead they are thinking about these in terms of sex-related characteristics like genitalia, chromosomes, or hormones.
- Emphasize how this interpretation does not leave space for **people who fall outside the gender/sex binary**, including those with nonbinary gender identities and intersex people.

2. You realize this crude analysis may not fully capture the relationship between respondent gender and depression. Draw a DAG that may more accurately capture the effect by considering the role of gender relations and/or sex-linked biology, as relevant. You may consider looking at Table 2 in this paper for inspiration, and the NHANES codebook to see what other variables are available to you.

TF notes:

- There are many possibilities here. The goal of this question is to get students thinking about how gender relations and sex-linked biology are **empirical questions that need to be considered in any analysis**, and how sex and/or gender are operationalized matters for doing good science.
- Consider emphasizing how sex and gender are likely **synergistic determinants** in relation to depression.
- Try to get students to explicate **what elements of sex and/or gender** are relevant for studying depression.
- Have students review the codebook to see **what is/is not possible** in terms of granularity of sex and/or gender variables. This helps emphasize the shortcomings of using secondary datasets for gender analysis.

Example DAG:


Social-cultural gender roles, including caregiving
(proxy: occupation)
Gendered stereotypes and expectations
(no variable in NHANES)
Social support and social relationships
(no variable in NHANES)
Norms around body image
(proxy: weight history)
Gender
Depression
Prenatal/postpartum/pubertal hormonal changes (proxy: FSH/LH)
Fluctuations in hormones across menstrual cycle
(proxy: estradiol)
Sex
Hormonal changes in transition to menopause
(proxy: estradiol, progesterone)
Age-related decline in hormones in males
(proxy: testosterone)

Inspiration drawn from Wierenga et al., 2023

4. You decide to re-run your analysis, now including occupation as a potential stratifying variable for the association between respondent gender and depression. Calculate the prevalence of reporting depression by gender separately by level of occupation. Comment on what you observe, including how this analysis leads to different conclusions compared to your answer in question 1.

```
### Question 3: re-run analysis, using occupational status

# recode occupation into working or looking for work vs. not working
nhanes$occ_2cat <- ifelse(nhanes$OCD150<=3, 1, 0)
nhanes$occ_2cat <- factor(nhanes$occ_2cat, levels=c(0,1), labels=c("no paid work",
"paid work or looking for paid work"))

# look at depression by occupation
table(nhanes$occ_2cat, nhanes$depr_2cat)

# create subsets of the data for each "gender"
female_only <- subset(nhanes, RIAGENDR=="female", select=c(1:7))
table(female_only$occ_2cat, female_only$depr_2cat)

male_only <- subset(nhanes, RIAGENDR=="male", select=c(1:7))
```

```
table(male_only$occ_2cat, male_only$depr_2cat)

# create subsets of the data for each occupational status
work_only <- subset(nhanes, occ_2cat=="paid work or looking for paid work",
select=c(1:7))
table(work_only$RIAGENDR, work_only$depr_2cat)

nowork_only <- subset(nhanes, occ_2cat=="no paid work", select=c(1:7))
table(nowork_only$RIAGENDR, nowork_only$depr_2cat)
```

Interpretation:
In the overall dataset, those with no paid work had 2.62 times the prevalence of moderate to severe depression compared to those with paid work or looking for paid work.

Among self-identified females, those with no paid work had 2.12 times the prevalence of moderate to severe depression compared to those with paid work or looking for paid work.

Among self-identified males, those with no paid work had 3.29 times the prevalence of moderate to severe depression compared to those with paid work or looking for paid work.

Among those with paid work or looking for paid work, respondents who self-identified as female had 2.14 times the prevalence of moderate to severe depression compared to survey respondents who self-identified as male.

Among those without paid work, respondents who self-identified as female had 1.37 times the prevalence of moderate to severe depression compared to survey respondents who self-identified as male.

Additional TF notes:

- Secondary data analysis is **limited to the information that was collected in the dataset**, and the extent of reporting regarding what each variable is capturing (i.e., what is included in the "gender" or "sex" variable(s). When working with secondary data, it is necessary to evaluate how these constructs are measured and reported.
- Asking what **dimensions of gender/sex are relevant** to the (causal) question of interest is critical in order to derive unbiased estimates of the parameter of interest.
- **Stratification** can be a useful approach to understand how associations of interest may differ by levels related to gender(ed) phenomena.
- Lack of exploration regarding differences by "gender" may result in the **wrong answer**. It may also perpetuate **misinformation and harmful stereotypes** about gender.

For future reading:
Christiansen, D. M., McCarthy, M. M., & Seeman, M. V. (2022). Understanding the influences of sex and gender differences in mental disorders. *Frontiers in Psychiatry*, *13*, 984195. doi: 10.3389/fpsyt.2022.984195

For more information about courses that emphasize gender/sex analysis:
https://www.hsph.harvard.edu/women-gender-and-health/info-sheet/