A Semester-long Class Project on the *All-of-Us* Database

HAP 819 – George Mason University

Guided by Prof. Alemi & Vladimir Franzuela Cardenas

Presented by Mohammad Qodrati

Aim and Materials

- To use medical history of patients to screen for a cancer

- The *All of Us* "Researcher Workbench" platform
  - The Registered Tier Dataset v7
  - The Cloud analysis environment

- R (programming language)

The All of Us Research Initiative

- Started in 2015
- Individualized medicine
- 1 million volunteers in the US
- Genetic and health data
- Sources:
  - Surveys on lifestyle & environment
  - EHR
  - Spot Labs and Physical measurements
  - Wearable devices

The Previously

Load datasets

```r
df_A <- readr::read_csv("./data/df_A.csv") # main demographics
df_B <- readr::read_csv("./data/df_B.csv") # has prostate cancer
df_C <- readr::read_csv("./data/df_C.csv") # has any diseases
df_D <- readr::read_csv("./data/df_D.csv") # has been observed with death
```

Datasets: A & B & D

```
head(df_A, 5)
```

A data.frame: 5 × 4

	person_id <dbl>	date_of_birth <chr>	race <chr>	ethnicity <chr>
1	[redacted]	1938-10-08 00:00:00 UTC	PMI: Skip	PMI: Skip
2	[redacted]	1944-09-14 00:00:00 UTC	PMI: Skip	PMI: Skip
3	[redacted]	1939-10-29 00:00:00 UTC	PMI: Skip	PMI: Skip
4	[redacted]	1999-08-20 00:00:00 UTC	PMI: Skip	PMI: Skip
5	[redacted]	1945-01-06 00:00:00 UTC	PMI: Skip	PMI: Skip

```
head(df_B, 5)
```

A data.frame: 5 × 3

	person_id <int>	standard_concept_code <dbl>	condition_start_datetime <chr>
1	[redacted]	4.36381e+14	2018-12-26 00:00:00 UTC
2	[redacted]	4.36381e+14	2020-05-18 00:00:00 UTC
3	[redacted]	2.54901e+08	2020-12-11 00:00:00 UTC
4	[redacted]	2.54901e+08	2018-07-22 12:01:00 UTC
5	[redacted]	2.54901e+08	2016-09-20 05:00:00 UTC

```
head(df_D, 5)
```

A data.frame: 5 × 2

	person_id <int>	observation_datetime <chr>
1	[redacted]	2021-04-14 04:26:00 UTC
2	[redacted]	2020-03-24 14:47:29 UTC
3	[redacted]	2021-09-15 23:40:30 UTC
4	[redacted]	2020-11-28 22:33:44 UTC
5	[redacted]	2022-05-21 00:49:15 UTC

- Date-time conversions
- Joins

Code(s) of the Outcome Disease(s)

Dataset C

- Contained all diseases.
- Should have been excluded:
  - The OUTCOME disease(s) (already in dataset B)
  - Diseases happened AFTER the outcome disease(s)

```
head(df_C, 5)
```

A data.frame: 5 × 4

	person_id <dbl>	standard_concept_name <chr>	standard_concept_code <dbl>	condition_start_datetime <chr>
1	[redacted]	Eustachian tube disorder	69494008	2019-08-02 06:00:00 UTC
2		Vitreous hemorrhage	31341008	2015-05-27 05:00:00 UTC
3		Recurrent major depressive episodes, moderate	191611001	2018-02-21 06:00:00 UTC
4		Low tension glaucoma	50485007	2016-04-09 00:00:00 UTC
5		Sepsis without septic shock	789043007	2022-03-23 06:00:00 UTC

Date-time Conversions

```r
# Change date_of_birth to a datetime format
df_A <- df_A %>%
  mutate(date_of_birth = lubridate::ymd_hms(date_of_birth))
```

```r
# Change concept codes to string instead of double which is shown with scientific notion
df_B$standard_concept_code <- as.character(df_B$standard_concept_code)

# Change condition_start_datetime to datetime format
df_B <- df_B %>%
  mutate(condition_start_datetime = lubridate::ymd_hms(condition_start_datetime))
```

```r
# Change concept codes to string instead of double which is shown with scientific notion
df_C$standard_concept_code <- as.character(df_C$standard_concept_code)

# Change condition_start_datetime to datetime format
df_C <- df_C %>%
  mutate(condition_start_datetime = lubridate::ymd_hms(condition_start_datetime))
```

```r
# Change observation_datetime to datetime format
df_D <- df_D %>%
  mutate(date_of_death = lubridate::ymd_hms(observation_datetime))
```

```r
# Process df_D to get only one date_of_death per patient

df_D_unique <- df_D %>%
  group_by(person_id) %>%
  summarize(date_of_death = min(observation_datetime, na.rm = TRUE)) %>%
```

```r
# Process df_B to get the date_of_first_diagnosis per patient

df_B_aggregated <- df_B %>%
  group_by(person_id) %>%
  summarize(date_of_first_diagnosis = min(condition_start_datetime, na.rm = TRUE)) %>%
  ungroup()
```

```r
df_final <- df_A
```

```r
df_final <- left_join(df_final, df_D_unique, by = "person_id")
```

```r
df_final <- left_join(df_final, df_B_aggregated, by = "person_id")
```

Social determinants of Health

```r
# The IDs of Persons with SDOH
patients_w_sdoh <- as.vector(read.csv('./data/df_persons_w_sdoh.csv'))
patients_w_sdoh <- patients_w_sdoh[[1]]
```

```r
df_analysis_w_sdoh <- df_final %>%
    rowwise() %>%
        mutate(sdoh  = if_else(person_id %in% patients_w_sdoh, 1, 0))
```


Exclude some diseases

- ```r
  # Create list of unique standard_concept_codes of the cancer of interest from df_B and store in df_exclude
  df_exclude <- df_B %>%
    distinct(standard_concept_code)

  # Remove rows from df_C if condition_concept_id is in df_exclude (exclude prostate cancer from double counting)
  df_C_filtered <- df_C %>%
    anti_join(df_exclude, by = "standard_concept_code")
  ```

- ```r
  # Remove rows in df_C where disease happened after cancer diagnosis
  df_C_filtered <- df_C_filtered %>%
    left_join(
        df_analysis %>% select(person_id, date_of_first_diagnosis), by = "person_id") %>%
    filter(
      is.na(date_of_first_diagnosis) | condition_start_datetime <= date_of_first_diagnosis
    ) %>%
    select(-(date_of_first_diagnosis))
  ```


Person-Outcome Data frame

```
df_person_vs_the_cancer <- df_analysis_w_sdoh %>%
    mutate(
        OUTCOME_Cancer = if_else(is.na(date_of_first_diagnosis), 0, 1)
    ) %>%
        select(c('person_id', 'OUTCOME_Cancer'))
```

Feature Construction

Grouping diseases into body systems using SNOMED-CT Hierarchy

Dataset of “Disease – Disease Group”s

```
df_disease_grouped <- read.csv("./data/df_disease_grouped.csv")
head(df_disease_grouped)
```

A data.frame: 6 × 2

	standard_concept_code	disease_group
	<dbl>	<int>
1	5.361900e+07	0
2	9.374501e+07	79604008
3	4.256710e+08	53619000
4	3.138401e+07	928000
5	7.741290e+08	928000
6	1.186352e+16	928000

362966006 - Disorder of auditory system
79604008 - Disorder of breast
49601007 - Disorder of cardiovascular system
53619000 - Disorder of digestive system
362969004 - Disorder of endocrine system
414027002 - Disorder of hematopoietic structure
414030009 - Disorder of immune structure
128598002 - Disorder of integument
362971004 - Disorder of lymphatic system
49483002 - Disorder of mediastinum
95351003 - Disorder of mucous membrane
928000 - Disorder of musculoskeletal system
118940003 - Disorder of nervous system
50043002 - Disorder of respiratory system
42030000 - Disorder of the genitourinary system

```
length(unique(df_disease_grouped$disease_group))
```

16

Add “Disease Group” column, …

```
df_C_grp <- df_C_filtered %>%
    # The group for each diagnosis is joined to the dataframe
    left_join(df_disease_grouped, by = "standard_concept_code")
```

	standard_concept_code	disease_group
	<dbl>	<int>
1	5.361900e+07	0
2	9.374501e+07	79604008
3	4.256710e+	53619000
4	3.138401 +07	928000
5	7.741 0e+08	928000
6	1 6352e+16	928000

A data.frame: 5 × 5

	person_id	standard_concept_name	standard_concept_code	condition_start_datetime	disease_group
	<dbl>	<chr>	<chr>	<dttm>	<int>
1		Eustachian tube disorder	69494008	2019-08-02 06:00:00	362966006
2		Vitreous hemorrhage	31341008	2015-05-27 05:00:00	928000
3		Recurrent major depressive episodes, moderate	191611001	2018-02-21 06:00:00	118940003
4		Low tension glaucoma	50485007	2016-04-09 00:00:00	118940003
5		Sepsis without septic shock	789043007	2022-03-23 06:00:00	0

… vs. altering the prev.-built column

```
# Add a new column called disease_group and initialize with NA_character_
# To be used later for further aggregation to create dummy variables
df_C_aggregated <- df_C_aggregated %>%
  mutate(disease_group = NA_character_)

# Create the list of disease groups we need
disease_groups <- df_disease_grouped$Code

# Iterate over each row of df_C_aggregated to update disease_group
df_C_aggregated <- df_C_aggregated %>%
  rowwise() %>%
  mutate(disease_group = ifelse(
    standard_concept_code %in% disease_groups,
    as.character(standard_concept_code),
    disease_group
  )) %>%
  ungroup()
```


```r
# Turn all 'character' variables into 'categorical (factors)' ones
df_C_grp_factorized <- df_C_grp
df_C_grp_factorized$disease_group <- as.factor(df_C_grp_factorized$disease_group)
df_C_grp_factorized$standard_concept_name <- as.factor(df_C_grp_factorized$standard_concept_name)
df_C_grp_factorized$standard_concept_code <- as.factor(df_C_grp_factorized$standard_concept_code)
```

```r
df_C_grp_named <- df_C_grp_factorized %>%
    mutate(disease_group =
               as.factor(case_match(disease_group,
                                    '0' ~ 'dz_others',
                                    '928000' ~ 'dz_MSK',
                                    '42030000' ~ 'dz_GU',
                                    '49483002' ~ 'dz_MDSTN',
                                    '49601007' ~ 'dz_CV',
                                    '50043002' ~ 'dz_Resp',
                                    '53619000' ~ 'dz_Digest',
                                    '79604008' ~ 'dz_BRST',
                                    '95351003' ~ 'dz_mucus',
                                    '118940003' ~ 'dz_NRVS',
                                    '128598002' ~ 'dz_Integ',
                                    '362966006' ~ 'dz_Audi',
                                    '362969004' ~ 'dz_ENCRN',
                                    '362971004' ~ 'dz_Lymph',
                                    '414027002' ~ 'dz_HematoP',
                                    '414030009' ~ 'dz_IMUN'
                           )))
```


Keep only 1 occurrence of each disease

```r
# Only one occurrnce of each disease per patient, without considering the times
df_one_of_each_dz <- df_C_grp_named %>%
    distinct(
        person_id,
        standard_concept_name,
        disease_group,
        .keep_all = TRUE
    )
```

or / versus

```r
# # The "First in time" Occurrence of each Disease
df_first_of_each_dz <- df_C_grp_named %>%
    select(person_id, standard_concept_name, disease_group)
        summarize(
            first_timer = min(condition_start_datetime),
            .by = c(person_id, standard_concept_name, disease_group)
        )
```


Add the Outcome column [and sort]

```
df_one_of_each_dz_with_OUTCOME <- df_one_of_each_dz %>%
    left_join(df_person_vs_the_cancer, by="person_id") %>%
    arrange(person_id,
            disease_group,
            standard_concept_name,
            .by_group = TRUE)
```

	person_id	standard_concept_name	standard_concept_code	condition_start_datetime	disease_group	OUTCOME_Cancer
	<chr>	<fct>	<fct>	<dttm>	<fct>	<dbl>
1		Anorectal disorder	426867001	2020-02-11 00:00:00	dz_Digest	0
2		Congenital anomaly of intestinal tract	126764002	2020-03-16 00:00:00	dz_Digest	0
3		Internal hemorrhoids grade II	721704005	2020-02-16 00:00:00	dz_Digest	0
4		Chronic prostatitis	19905009	2020-01-04 00:00:00	dz_GU	0
5		Hematuria syndrome	53298000	2020-06-06 00:00:00	dz_GU	0
6		Microscopic hematuria	197940006	2020-07-14 00:00:00	dz_GU	0
7		Diaphragmatic hernia	39839004	2020-03-16 00:00:00	dz_MSK	0
8		Organic mental disorder	111479008	2012-07-22 05:00:00	dz_NRVS	0

Drop the date-time column

```
df_one_dz_grp_OUTCOME_no_time <- select(df_one_of_each_dz_with_OUTCOME,
                                       -c('condition_start_datetime'))
```

Associations:
Outcome (+/-) ↔ Exposure (+/-)

For each body system, estimate the association of the cancer with the history of the conditions falling within that body system.

Methods to this Aim

1. Calculation of Likelihood Ratios using the formula:

$$LR = \frac{\text{Probability of finding in patients } \textit{with} \text{ disease}}{\text{Probability of the same finding in patients } \textit{without} \text{ disease}}$$

2. Fitting Statistical Models (Regression)

3. …

Likelihood Ratio

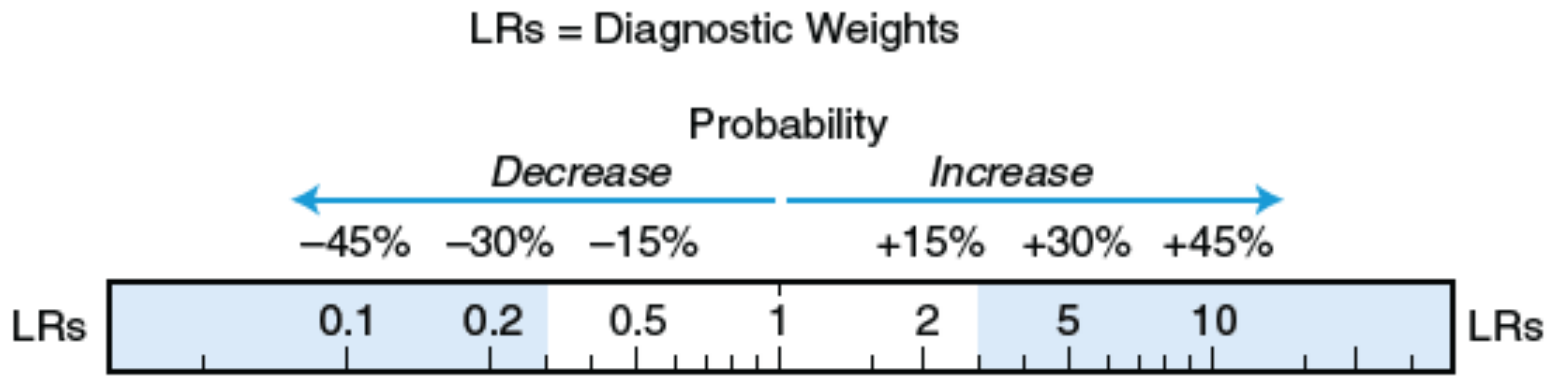


FIG. 2.5 APPROXIMATING PROBABILITY. Clinicians can estimate changes in probability by recalling the LRs 2, 5, and 10 and the first 3 multiples of 15 (i.e., 15, 30, and 45). A finding whose LR is 2 increases probability about 15%, one of 5 increases it 30%, and one of 10 increases it 45% (these changes are *absolute* increases in probability). LRs whose values are 0.5, 0.2, and 0.1 (i.e., the reciprocals of 2, 5, and 10) decrease probability 15%, 30%, and 45%, respectively. Throughout this book, LRs with values ≥3 or ≤0.3 (represented by the shaded part of the diagnostic weight "ruler") are presented in boldface type to indicate those physical findings that change probability sufficiently to be clinically meaningful (i.e., they increase or decrease probability at least 20% to 25%).

National Institutes of Health (
https://pubmed.ncbi.nlm.nih.gov ›

Does This Patient With

by AC Fanaroff · 2015 · Cited by 27

pain are ultimately diagnosed with a

Table 1. Performance of Cardiac Risk Factors in Diagnosing Acute Coronary Syndrome[a]

	No.		% (95% CI)						%[b]	
Test	Studies	Patients	Sensitivity	Specificity	LR+ (95% CI)	I^2, %	LR− (95% CI)	I^2, %	PPV	NPV
Abnormal prior stress[c,61]	1	1777	12 (8-16)	96 (95-97)	3.1 (2.0-4.7)		0.92 (0.88-0.96)		32	12
Peripheral arterial disease[21,23,49]	3	6034	7.5 (2-11)	97 (95-99)	2.7 (1.5-4.8)	0	0.96 (0.94-0.98)	64	29	13
Prior CAD[37,40,49,57,60]	5	6396	41 (13-69)	79 (60-98)	2.0 (1.4-2.6)	87	0.75 (0.56-0.93)	96	23	10
Prior myocardial infarction[d]	9	10 491	28 (21-36)	82 (78-86)	1.6 (1.4-1.7)	42	0.88 (0.81-0.93)	81	19	12
Diabetes[e]	9	10 237	26 (21-32)	82 (77-85)	1.4 (1.3-1.6)	4	0.90 (0.86-0.94)	45	17	12
Cerebrovascular disease[21,23,49,70]	4	6682	10 (8-13)	93 (91-94)	1.4 (1.1-1.8)	18	0.97 (0.94-0.99)	14	17	13
Men[f]	12	21 113	66 (62-76)	50 (44-51)	1.3 (1.2-1.3)	65	0.70 (0.64-0.77)	39	16	9
Hyperlipidemia[g]	10	10 288	42 (31-55)	67 (56-79)	1.3 (1.1-1.5)	70	0.85 (0.77-0.93)	69	16	11
Hypertension[h]	11	10 931	59 (53-66)	52 (44-60)	1.2 (1.1-1.3)	51	0.78 (0.72-0.85)	29	15	10
Any tobacco use[i]	9	7 381	38 (28-47)	65 (55-75)	1.1 (0.9-1.3)	75	0.96 (0.85-1.1)	77	14	13
Family history of CAD[21,23,40,49,51,54,58]	7	8 717	37 (26-47)	64 (58-71)	1.0 (0.9-1.2)	54	0.99 (0.91-1.1)	65	13	13
Obesity[21,41,60]	3	4887	40 (26-55)	68 (48-84)	1.0 (0.9-1.2)	45	0.99 (0.88-1.1)	44	13	13
Prior CABG[23,31,58,70]	4	5902	9.1 (6-14)	91 (87-94)	0.97 (0.5-2.1)	77	1.00 (0.92-1.1)	77	13	13

Abbreviations: CABG, coronary artery bypass graft; CAD, coronary artery disease; LR+, positive likelihood ratio; LR−, negative likelihood ratio; NPV, negative predictive value; PPV, positive predictive value.

[a] See eTable 4 in the Supplement for results from individual studies.

[b] PPV and NPV calculated assuming an acute coronary syndrome rate of 13%. The included studies had an acute coronary syndrome rate of 13% (95% CI, 11%-16%).

[c] When the summary measure was from less than 3 studies, the I^2 was not calculated.

[d] References 21, 23, 37, 49, 54, 58, 60, 70.

[e] References 21, 23, 31, 40, 49, 51, 58, 62, 70.

[f] References 21, 23, 31, 40, 47, 49, 51, 54, 58, 60, 62, 70.

[g] References 21, 23, 40, 49, 51, 54, 58, 60, 62, 70.

[h] References 21, 23, 31, 40, 49, 51, 54, 58, 60, 62, 70.

[i] References 21, 31, 40, 49, 51, 54, 58, 60, 62.

Pleuritic pain[e,37,49]	2	3487	18-36	78-93	0.35-0.61		1.1-1.2		6.6-8.4	14-15

Method 1: Calculation of Likelihood Ratios

#

standard_concept_name	disease_group	OUTCOME_Cancer	dx_cnt
<chr>	<chr>	<dbl>	<dbl>
Ablepharon	dz_Audi	0	1
Ablepharon	dz_Audi	1	1
Abnormal auditory perception	dz_Audi	0	640
Abnormal auditory perception	dz_Audi	1	36
Abscess of external auditory canal	dz_Audi	0	1
Abscess of external ear	dz_Audi	0	28

```
df
                                                    ncer)
                                                    Cancer)
```

Risk or exposure factor	Disease	Non disease	
Exposure	a	b	a+b
Non exposure	c	d	c+d
	a+c	b+d	a+b+c+d

Having them side-by-side

standard_concept_name <chr>	disease_group <chr>	OUTCOME_Cancer <dbl>	dx_cnt <dbl>
Ablepharon	dz_Audi	0	1
Ablepharon	dz_Audi	1	1
Abnormal auditory perception	dz_Audi	0	640
Abnormal auditory perception	dz_Audi	1	36
Abscess of external auditory canal	dz_Audi	0	1
Abscess of external ear	dz_Audi	0	28

```
df_outcome_pos_dz_counts <- df_dz_counts[df_dz_counts$OUTCOME_Cancer == 1,
                                        c('standard_concept_name',
                                          'disease_group',
                                          'dx_cnt'
                                         )]
names(df_outcome_pos_dz_counts)[3] <- 'p_dx_cnt'

df_outcome_neg_dz_counts <- df_dz_counts[df_dz_counts$OUTCOME_Cancer == 0,
                                        c('standard_concept_name',
                                          'disease_group',
                                          'dx_cnt'
                                         )]
names(df_outcome_neg_dz_counts)[3] <- 'n_dx_cnt'
```

```
df_outcome_dz_counts_joined <- df_outcome_pos_dz_counts %>%
    full_join(df_outcome_neg_dz_counts,
          by=c('standard_concept_name','disease_group'))
```


standard_concept_name	disease_group	OUTCOME_Cancer	dx_cnt
<chr>	<chr>	<dbl>	<dbl>
Ablepharon	dz_Audi	0	1
Ablepharon	dz_Audi	1	1
Abnormal auditory perception	dz_Audi	0	640
Abnormal auditory perception	dz_Audi	1	36
Abscess of external auditory canal	dz_Audi	0	1
Abscess of external ear	dz_Audi	0	28

standard_concept_name	disease_group	p_dx_cnt	n_dx_cnt
<chr>	<chr>	<dbl>	<dbl>
Ablepharon	dz_Audi	1	1
Abnormal auditory perception	dz_Audi	36	640
Acquired stenosis of external ear canal	dz_Audi	3	40
Active cochlear Ménière's disease	dz_Audi	4	15
Active cochleovestibular Ménière's disease	dz_Audi	1	21
Active vestibular Ménière's disease	dz_Audi	1	4
Acute actinic otitis externa	dz_Audi	1	32
Acute contact otitis externa	dz_Audi	2	34

Non-existent “a / b” s

Wound of sternal region	dz_others	NA	1
X-linked hereditary disease	dz_others	NA	1
Yaws	dz_others	NA	7
Yaws gummata and ulcers	dz_others	NA	2
Yellow fever	dz_others	NA	5
Zinc deficiency	dz_others	NA	3
Zinc poisoning	dz_others	NA	1
Zygomycosis	dz_others	NA	8

```
df_for_LRs <- df_outcome_dz_counts_joined %>%
    replace_na(list(p_dx_cnt = 0, n_dx_cnt = 0))
```

LRs

Risk or exposure factor	Disease	Non disease	
Exposure	a	b	a+b
Non exposure	c	d	c+d
	a+c	b+d	a+b+c+d

```
n_total_outcome_pos <- sum(df_person_vs_the_cancer$OUTCOME_Cancer)
n_total_outcome_neg <- length(df_person_vs_the_cancer$OUTCOME_Cancer) - n_total_outcome_pos

cat('Total # of people with+++ the cancer of interest:', n_total_outcome_pos)
cat('\nTotal # of people without the cancer of interest:', n_total_outcome_neg)
```

```
Total # of people with+++ the cancer of interest: 6122
Total # of people without the cancer of interest: 147460
```

a + c
b + d

```
df_with_LRs <- df_for_LRs %>%
    mutate(
        pLR <- case_when(
            n_dx_cnt == 0 ~ NA_real_, # zero denominator = can be addressed with considering only one subject
            p_dx_cnt == 0 ~ 0, # zero denominator
            .default = (p_dx_cnt/n_total_outcome_pos) / (n_dx_cnt/n_total_outcome_neg)
        )
    )
```

```
names(df_with_LRs)[5] <- 'pLR'
```

standard_concept_name	disease_group	p_dx_cnt	n_dx_cnt	pLR
<chr>	<chr>	<dbl>	<dbl>	<dbl>
Ablepharon	dz_Audi	1	1	24.0868997
Atrophic nonflaccid tympanic membrane	dz_Audi	1	45	0.5352644
Bannayan syndrome	dz_Audi	1	0	NA
Benign paroxysmal positional vertigo	dz_Audi	114	1945	1.4117772
Benign tumor of external ear	dz_Audi	2	9	5.3526444
Bilateral disorder of ears	dz_Audi	3	61	1.1846016
Bilateral disorder of mastoids	dz_Audi	1	3	8.0289666
⋮	⋮	⋮	⋮	⋮
Visceral leishmaniasis	dz_others	0	1	0
Visual agnosia	dz_others	0	8	0
Vitamin A deficiency with corneal ulceration AND xerosis	dz_others	0	1	0
Vitamin A deficiency with ocular manifestation	dz_others	0	10	0
Vitamin B12 deficiency (non anemic)	dz_others	0	4	0
Vitamin B12 deficiency anemia due to malabsorption with proteinuria	dz_others	0	14	0
Vitamin C deficiency anemia	dz_others	0	1	0
Vitamin E deficiency	dz_others	0	15	0
Vitamin K deficiency	dz_others	0	42	0
Vortex keratopathy	dz_others	0	1	0
Vortex keratopathy of bilateral eyes	dz_others	0	1	0

Method 2: Calculation of Odds Ratios

Split and Exclude

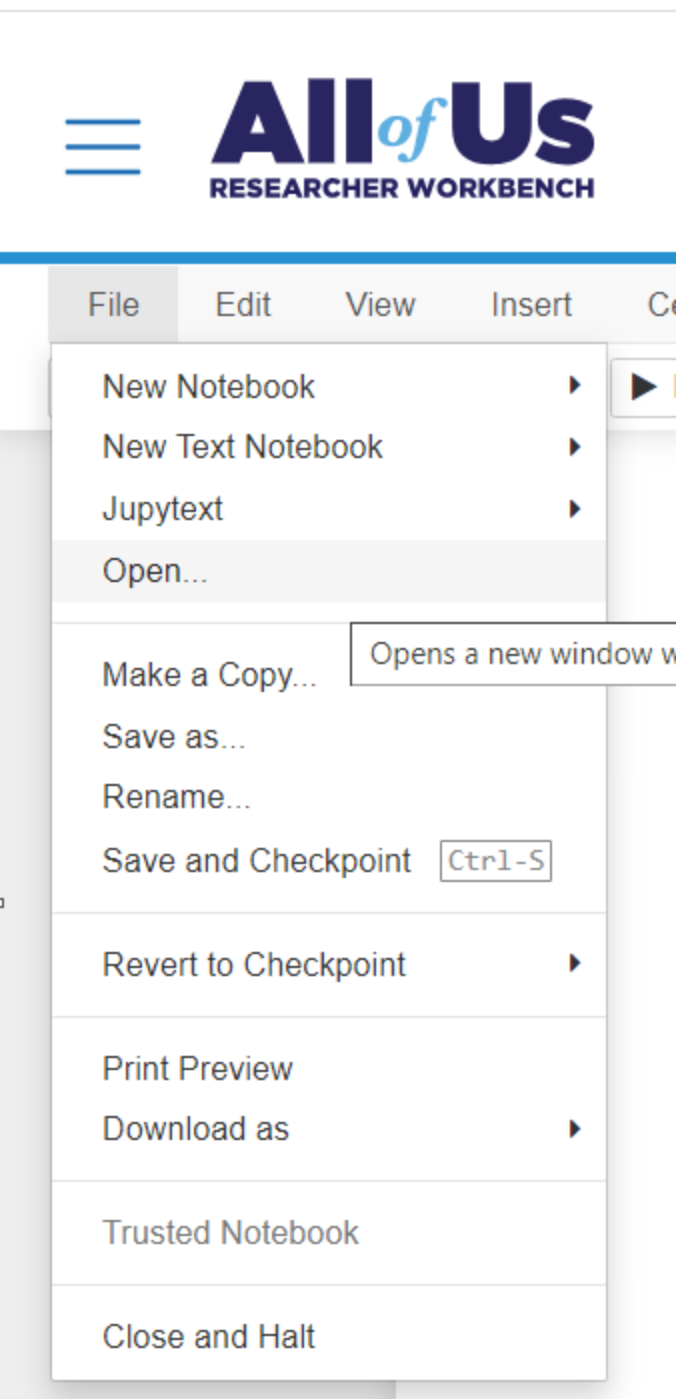
- Split based on the value of “disease group”:

```
list_df_based_on_grp <- df_one_dz_grp_OUTCOME_no_time %>%
    split(f=df_one_dz_grp_OUTCOME_no_time$disease_group)
```

```
for (df_grp in list_df_based_on_grp) {
    write_csv(df_grp,
              paste0('./data/new_ready_for_analysis/df_', unique(df_grp$disease_group), '_all.csv'))
}
```

- Keep only those conditions with >30 occurrence:

```
for (df_grp in list_df_based_on_grp) {
    write_csv(
        df_grp %>%
            group_by(standard_concept_name) %>%
            filter(n() > 30),
              paste0('./data/new_ready_for_analysis/df_', unique(df_grp$disease_group), '_gt30.csv'))
}
```


All of Us
RESEARCHER WORKBENCH
File
Edit
View
Insert
New Notebook
New Text Notebook
Jupytext
Open...
Opens a new window w
Make a Copy...
Save as...
Rename...
Save and Checkpoint
Ctrl-S
Revert to Checkpoint
Print Preview
Download as
Trusted Notebook
Close and Halt


df_dz_Audi_all.csv
df_dz_Audi_gt30.csv
df_dz_BRST_all.csv
df_dz_BRST_gt30.csv
df_dz_CV_all.csv
df_dz_CV_gt30.csv
df_dz_Digest_all.csv
df_dz_Digest_gt30.csv
df_dz_ENCRN_all.csv
df_dz_ENCRN_gt30.csv

Binary diseases, aggregated to just a row for a patient

```r
bs_load_dum_aggr_save <- function(body_system) {
    df_gt30 <- read_csv(paste0('./data/new_ready_for_analysis/df_dz_', body_system, '_gt30.csv'))

    print(body_system)

    df_gt30_dum_aggr <- bind_cols(
        select(df_gt30, c(1,5)),
        model.matrix(~standard_concept_name - 1, data = df_gt30)
    ) %>%
        summarize(across(everything(),max), .by=person_id)

    print(table(df_gt30_dum_aggr$OUTCOME_Cancer))

    write_csv(df_gt30_dum_aggr,
              paste0('./data/new_ready_for_analysis/df_', body_system, '_gt30_dum_aggr.csv'))

    return(df_gt30_dum_aggr)
}
```

Example: Disorders of Hematopoiesis

tandard_concept_nameCommon variable agammaglobulinemia	standard_concept_nameDisorder characterized by eosinophilia	standard_concept_nameDrug-induced immunodeficiency	standard_concept_nameDrug-induced neutropenia	standard_concept_nameEosinophilic asthma	...	standard_
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	...	
0	0	0	1	0	...	
0	0	1	0	0	...	
0	0	0	0	1	...	
0	0	0	0	0	...	
0	0	0	0	0	...	
0	0	0	0	0	...	

Regress the outcome on diseases in each group

```r
read_glm <- function(body_system) {
    df_dz_grp <- read_csv(paste0('./data/new_ready_for_analysis/df_', body_system, '_gt30_dum_aggr.csv'))
    model <- glm(OUT
    print(with(summa
    return(summary(m
}
```

```r
read_glm('HematoP')
```

```
Rows: 1863 Columns: 31
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (31): person_id, OUTCOME_Cancer, standard_concept_nameAcute leukemia, st...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Warning message:
“glm.fit: fitted probabilities numerically 0 or 1 occurred”
[1] 0.08971026


Call:
glm(formula = OUTCOME_Cancer ~ ., family = "binomial", data = select(df_dz_grp,
    -c("person_id")))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2037  -0.3687  -0.3308  -0.1979   3.2057
```

	Estima	Std. Err	z value	Pr(>|z|)		OR	pLR
Chronic lymphoid leukemia, disease	1.74476	0.34441	5.066	4.07E-07	***	5.724527424	3.679943010635595
B-cell chronic lymphocytic leukemia	-0.38172	0.36774	-1.038	0.2993		0.682686179	2.0605312432842915
Chronic lymphoid leukemia in remission	-0.09389	0.41978	-0.224	0.823		0.910382899	2.0072416421648702
Thalassemia	0.64707	0.42479	1.523	0.1277		1.909936509	1.952991868052306
Acute myeloid leukemia in remission	0.6422	0.72383	0.887	0.375		1.90065773	1.8178792230927123
Leukemia in remission	0.2938	0.67845	0.433	0.665		1.341515574	1.7624560760472026
Leukemia	0.30905	0.37078	0.834	0.4046		1.362130475	1.7204928361413168
Polycythemia vera (clinical)	0.11232	0.32355	0.347	0.7285		1.118870842	1.6984352356779668
Acute myeloid leukemia, disease	0.47944	0.39446	1.215	0.2242		1.615169654	1.5976004907026518
Acute lymphoid leukemia in remission	0.10029	0.60819	0.165	0.869		1.105491464	1.5640843964921063
Acute lymphoid leukemia	-0.05448	0.48164	-0.113	0.9099		0.946977448	1.5293269654589485
Congenital anomaly of spleen	-0.05081	0.56461	-0.09	0.9283		0.950459241	1.5293269654589485
Aplastic anemia	-0.1116	0.41179	-0.271	0.7864		0.894401947	1.3807776901516302
Disorder of transplanted bone marrow	-0.23545	0.55244	-0.426	0.67		0.790215173	1.353196612695418
Congenital absence of spleen	0.03866	0.63586	0.061	0.9515		1.039417022	1.2677315634725492
Lymphoid leukemia	-1.315	0.41433	-3.174	0.0015	**	0.268474323	1.1004167378365783
Acute leukemia	-0.59097	0.82868	-0.713	0.4758		0.553789848	0.9445843021952328
Heterozygous thalassemia	-0.62764	0.55848	-1.124	0.2611		0.533850202	0.8758872620355795
Myelofibrosis	-0.56618	0.73986	-0.765	0.4441		0.567689877	0.8758872620355795
Chronic myeloid leukemia	-0.74744	0.67796	-1.102	0.2703		0.47357736	0.5827475735317363
Sickle cell-hemoglobin C disease without crisis	0.50592	1.12449	0.45	0.6528		1.658510649	0.5601604582785683
Acute leukemia in remission	-1.49367	0.85408	-1.749	0.0803	.	0.224547054	0.5179978431393212
Beta thalassemia	-1.15619	1.04301	-1.109	0.2676		0.314682841	0.4817379941195688
Chronic myeloid leukemia in remission	-1.11187	1.18389	-0.939	0.3476		0.328943262	0.43794363101778977
Myeloid leukemia	-1.00903	1.0793	-0.935	0.3498		0.364572443	0.35950596576087224
Hemoglobin SS disease with crisis	-0.34356	1.1509	-0.299	0.7653		0.709240925	0.2483185536698 8081
Hemoglobin SS disease without crisis	-2.00995	1.10429	-1.82	0.0687	.	0.133995374	0.15341974335018113
Glucose-6-phosphate dehydrogenase deficiency anemia	-14.7063	650.685	-0.023	0.982		4.10319E-07	0
Sickle cell-hemoglobin SS disease	-12.9694	637.538	-0.02	0.9838		2.33063E-06	0
(Intercept)	-2.76671	0.24997	-11.068	< 2e-16	***	0.062868502	

Sorted by LRs

It’s easy to lie with statistics. It’s hard to tell the truth without statistics.

Andrejs Dunkels

coresignal®