Session 5 notes

Tim Riffe

7/30/2021

Summary

1. Download the data
2. Harmonize the data
3. Join the data
4. Visualize it

GBD

```
library(here)
```

```
## here() starts at /home/tim/workspace/KOSTAT_Workshop1
```

```
gbd_url <- "https://s3.healthdata.org/gbd-api-2017-public/835b25c27d7b31e221f6c51f7756875b_files/IHME-GI

local_file_gbd <- here("Data","GBD_prevalence.zip")

if (!file.exists(local_file_gbd)){
  download.file(gbd_url, destfile = local_file_gbd)
}
```

HLD

```
hld_url <- "https://www.lifetable.de/data/hld.zip"
local_file_hld <- here("Data","HLD.zip")

if (! file.exists(local_file_hld)){
  download.file(hld_url, destfile = local_file_hld)
}
```

Read in the data

GBD

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
GBD <- read_csv(local_file_gbd)
```

```
## Multiple files in zip: reading 'IHME-GBD_2017_DATA-835b25c2-1.csv'
```

```
## Rows: 49140 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): measure, location, sex, age, cause, metric
## dbl (4): year, val, upper, lower
```

```
## 
## 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.
```

```
glimpse(GBD)
```

```
## Rows: 49,140
## Columns: 10
## $ measure  <chr> "Prevalence", "Prevalence", "Prevalence", "Prevalence", "Prev~
## $ location <chr> "Ghana", "Ghana", "Ghana", "Ghana", "Ghana", "Ghana", "Ghana"~
## $ sex      <chr> "Male", "Female", "Both", "Male", "Female", "Both", "Male", "~
## $ age      <chr> "1 to 4", "1 to 4", "1 to 4", "5 to 9", "5 to 9", "5 to 9", "~
## $ cause    <chr> "Diabetes and kidney diseases", "Diabetes and kidney diseases~
## $ metric   <chr> "Percent", "Percent", "Percent", "Percent", "Percent", "Perce~
## $ year     <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2~
## $ val      <dbl> 0.0006772738, 0.0013323786, 0.0010007344, 0.0037088414, 0.004~
## $ upper    <dbl> 0.001216079, 0.002191553, 0.001592305, 0.005200965, 0.0065402~
## $ lower    <dbl> 0.0002696305, 0.0007487865, 0.0005419410, 0.0024902408, 0.003~
```

```
GBD$cause %>% unique()
```

```
## [1] "Diabetes and kidney diseases"   "Digestive diseases"
## [3] "Cardiomyopathy and myocarditis" "All causes"
```

HLD

```
HLD <- read_csv(local_file_hld)
```

```
## Rows: 1613643 Columns: 21
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country, Region, Residence, Ethnicity
## dbl (17): SocDem, Version, Ref-ID, Year1, Year2, TypeLT, Sex, Age, AgeInt, m...
```

```
## 
## 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.
```

Harmonize the data to be able to merge

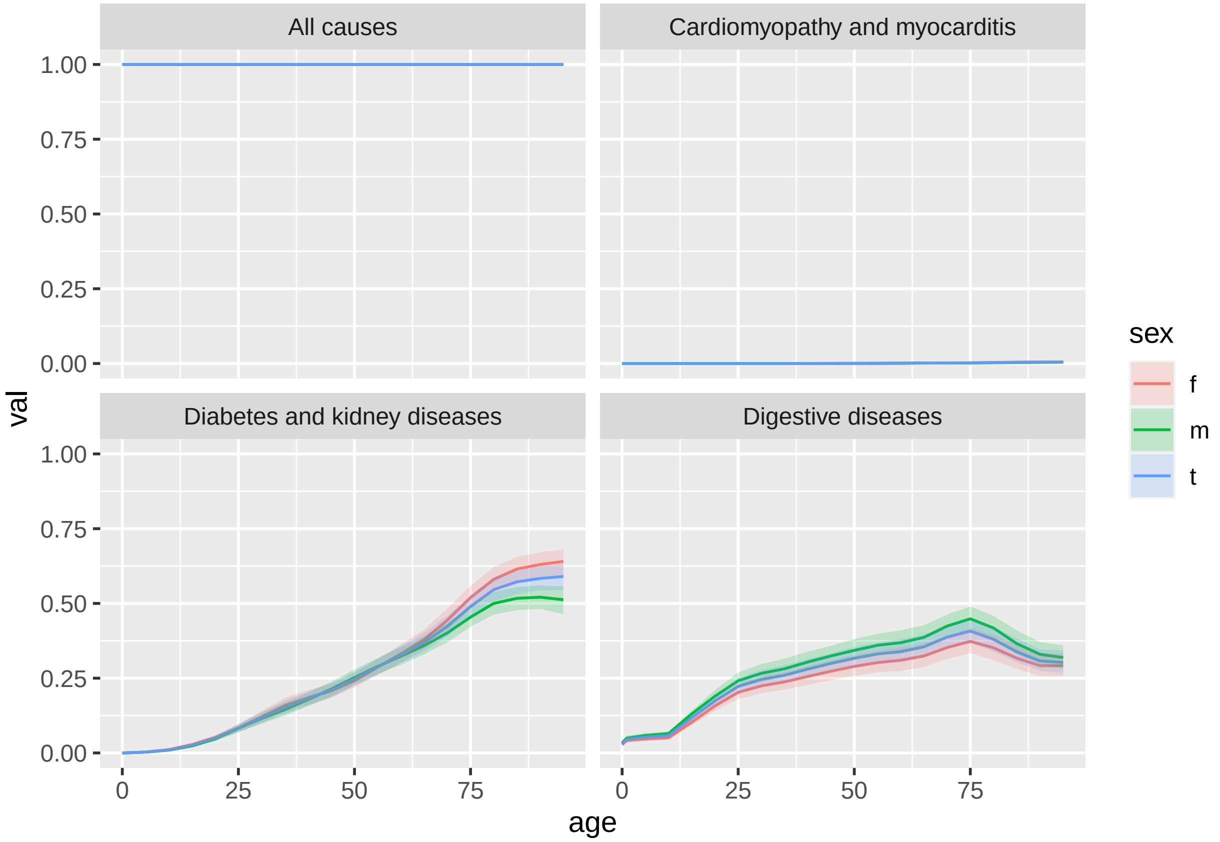
GBD

```
# install.packages("countrycode")
library(countrycode)
GBD$age %>% unique()
```

```
##  [1] "1 to 4"   "5 to 9"   "10 to 14" "15 to 19" "20 to 24" "25 to 29"
##  [7] "30 to 34" "35 to 39" "40 to 44" "45 to 49" "50 to 54" "55 to 59"
## [13] "60 to 64" "65 to 69" "70 to 74" "75 to 79" "<1 year"  "80 to 84"
## [19] "85 to 89" "90 to 94" "95 plus"
```

```
# substr("My name is Tim", start = 1, stop = 2)
GBD <-
  # incoming GBD, will only work on first run, because
  # we over-write
  GBD %>%
         # recode age in 3 steps
  mutate(age = substr(age, start = 1, stop = 2),
         age = ifelse(age == "<1", "0", age),
         age = parse_number(age),
         # recode sex to some standard
         sex = case_when(sex == "Male" ~ "m",
                         sex == "Female" ~ "f",
                         sex == "Both" ~ "t"),
         # use the countrycode package to find ISO3 codes
         # because that's easier to match on.
         ISO3 = countrycode(location,
                             origin = "country.name",
                             destination = "iso3c"))
```

```
GBD %>%
  filter(location == "India") %>%
  ggplot(aes(x = age, y = val, color = sex)) +
  geom_line() +
  geom_ribbon(aes(ymin = lower,
                  ymax = upper,
                  fill = sex),
              alpha = .2,
              color = NA)+
  facet_wrap(~cause)
```



Filter down

```
GBD <-
  GBD %>%
  filter(cause != "All causes")
```

HLD prep

```
HLD <-
  HLD %>%
  # filter down to nationally representative lifetables
  filter(Ethnicity == "0",
         Residence == "0",
         Region == "0",
         # abridged ages are codes > 1
         TypeLT > 1,
         # recent years only
         Year1 >= 2012) %>%

  # create a distance to 2017.5 indicator
  mutate(Year = (Year1 + Year2 + 1) / 2,
         Diff = 2017.5 - Year,
         Dist = abs(Diff)) %>%

  # figure out which subsets minimize this
  group_by(Country, Sex) %>%
  mutate(keep = Dist == min(Dist)) %>%

  # create a second stricter condition, just in case
  # we have equal distances on the left and right
  group_by(Country, Sex, keep) %>%
  mutate(keep2 = keep & Year == max(Year)) %>%
  ungroup() %>%
  filter(keep2) %>%
  # group down the open age to the lowest
  # common denominator. No effect on LE
  mutate(Age = ifelse(Age >= 80, 80, Age)) %>%
  group_by(Country, Sex, Age) %>%
  summarize(nLx = sum(`L(x)`),
            .groups = "drop") %>%
  # harmonize radix to 1
  mutate(radix = ifelse(Country %in% c("Malta","Turkmenistan"), 10000, 100000),
         nLx = nLx / radix,
         Sex = ifelse(Sex == 1, "m","f")) %>%

  # take care of renaming when we select final columns
  select(ISO3 = Country,
         sex = Sex,
         age = Age,
         nLx)
```

Ready to join

```
head(GBD)
```

```
## # A tibble: 6 x 11
##   measure  location sex     age cause metric  year      val   upper   lower ISO3
##   <chr>    <chr>    <chr> <dbl> <chr> <chr>  <dbl>    <dbl>   <dbl>   <dbl> <chr>
## 1 Prevale~ Ghana    m         1 Diab~ Perce~  2017 6.77e-4 0.00122 2.70e-4 GHA
## 2 Prevale~ Ghana    f         1 Diab~ Perce~  2017 1.33e-3 0.00219 7.49e-4 GHA
## 3 Prevale~ Ghana    t         1 Diab~ Perce~  2017 1.00e-3 0.00159 5.42e-4 GHA
## 4 Prevale~ Ghana    m         5 Diab~ Perce~  2017 3.71e-3 0.00520 2.49e-3 GHA
## 5 Prevale~ Ghana    f         5 Diab~ Perce~  2017 4.81e-3 0.00654 3.33e-3 GHA
## 6 Prevale~ Ghana    t         5 Diab~ Perce~  2017 4.25e-3 0.00570 3.05e-3 GHA
```

```
head(HLD)
```

```
## # A tibble: 6 x 4
##   ISO3  sex     age   nLx
##   <chr> <chr> <dbl> <dbl>
## 1 AUS   m         0 0.997
## 2 AUS   m         1 3.98
## 3 AUS   m         5 4.98
## 4 AUS   m        10 4.98
## 5 AUS   m        15 4.97
## 6 AUS   m        20 4.96
```

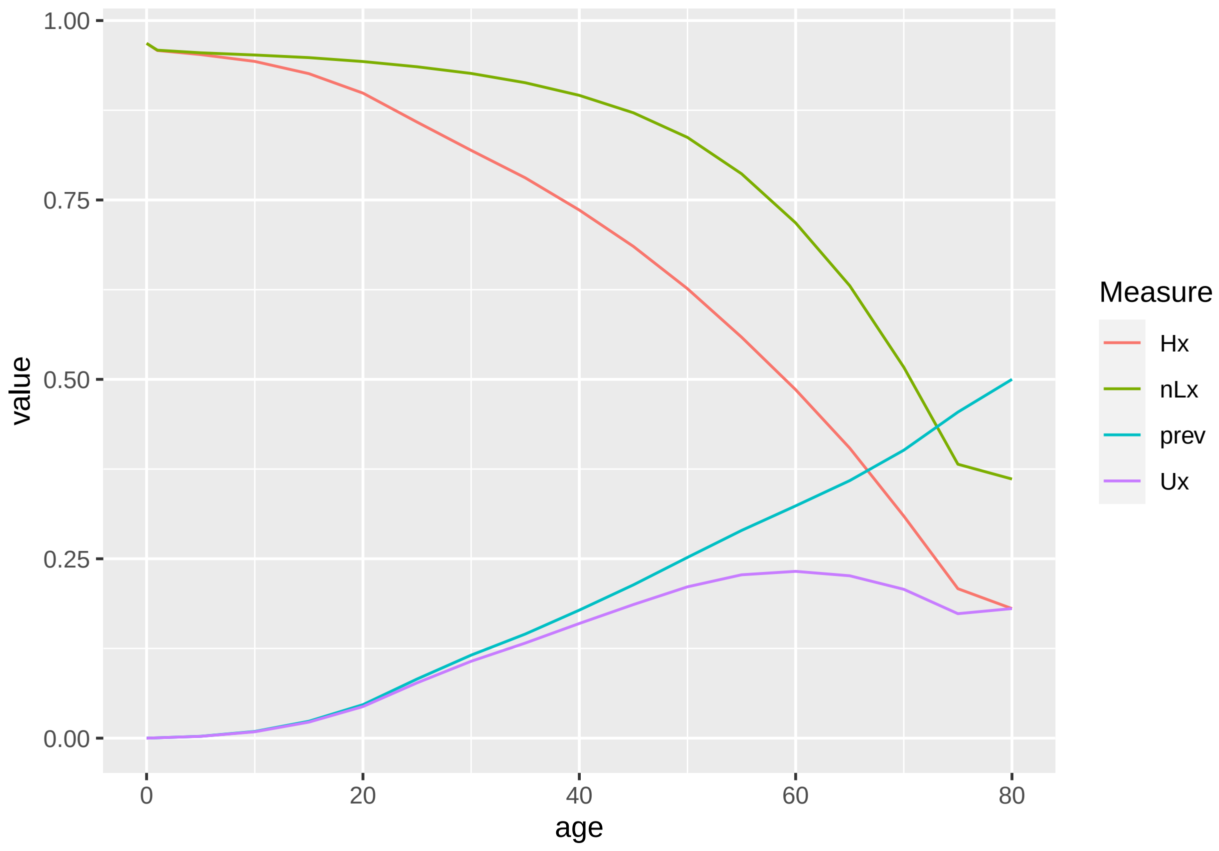
```
HLE <-
  inner_join(GBD, HLD, by = c("ISO3", "sex","age"))
```

```
HLE$location %>% unique() %>% sort()
```

```
##  [1] "Algeria"                    "Australia"
##  [3] "Austria"                    "Bangladesh"
##  [5] "Belgium"                    "Bosnia and Herzegovina"
##  [7] "Botswana"                   "Brazil"
##  [9] "Bulgaria"                   "Canada"
## [11] "Costa Rica"                 "Cyprus"
## [13] "Czech Republic"             "Denmark"
## [15] "Estonia"                    "Finland"
## [17] "France"                     "Georgia"
## [19] "Germany"                    "Hungary"
## [21] "Iceland"                    "India"
## [23] "Israel"                     "Italy"
## [25] "Japan"                      "Kazakhstan"
## [27] "Latvia"                     "Luxembourg"
## [29] "Macedonia"                  "Malaysia"
## [31] "Malta"                      "Mauritius"
## [33] "Netherlands"                "New Zealand"
## [35] "Norway"                     "Poland"
## [37] "Portugal"                   "Russian Federation"
## [39] "Serbia"                     "Singapore"
## [41] "Slovakia"                   "Slovenia"
## [43] "South Korea"                "Spain"
## [45] "Sweden"                     "Switzerland"
## [47] "Taiwan (Province of China)" "Tajikistan"
## [49] "Turkey"                     "United Kingdom"
## [51] "United States"
```

Calculate HLE

```
HLE %>%
  filter(location == "India",
         cause == "Diabetes and kidney diseases",
         sex == "m") %>%
  mutate(n = case_when(age == 0 ~ 1,
                       age == 1 ~ 4,
                       TRUE ~ 5),
         nLx = nLx / n,
         Ux = nLx * val,
         Hx = nLx - Ux) %>%
  select(age, nLx, prev = val, Ux, Hx) %>%
  pivot_longer(nLx:Hx,
               names_to = "Measure",
               values_to = "value") %>%
  ggplot(aes(x = age, y = value, color = Measure)) +
  geom_line()
```



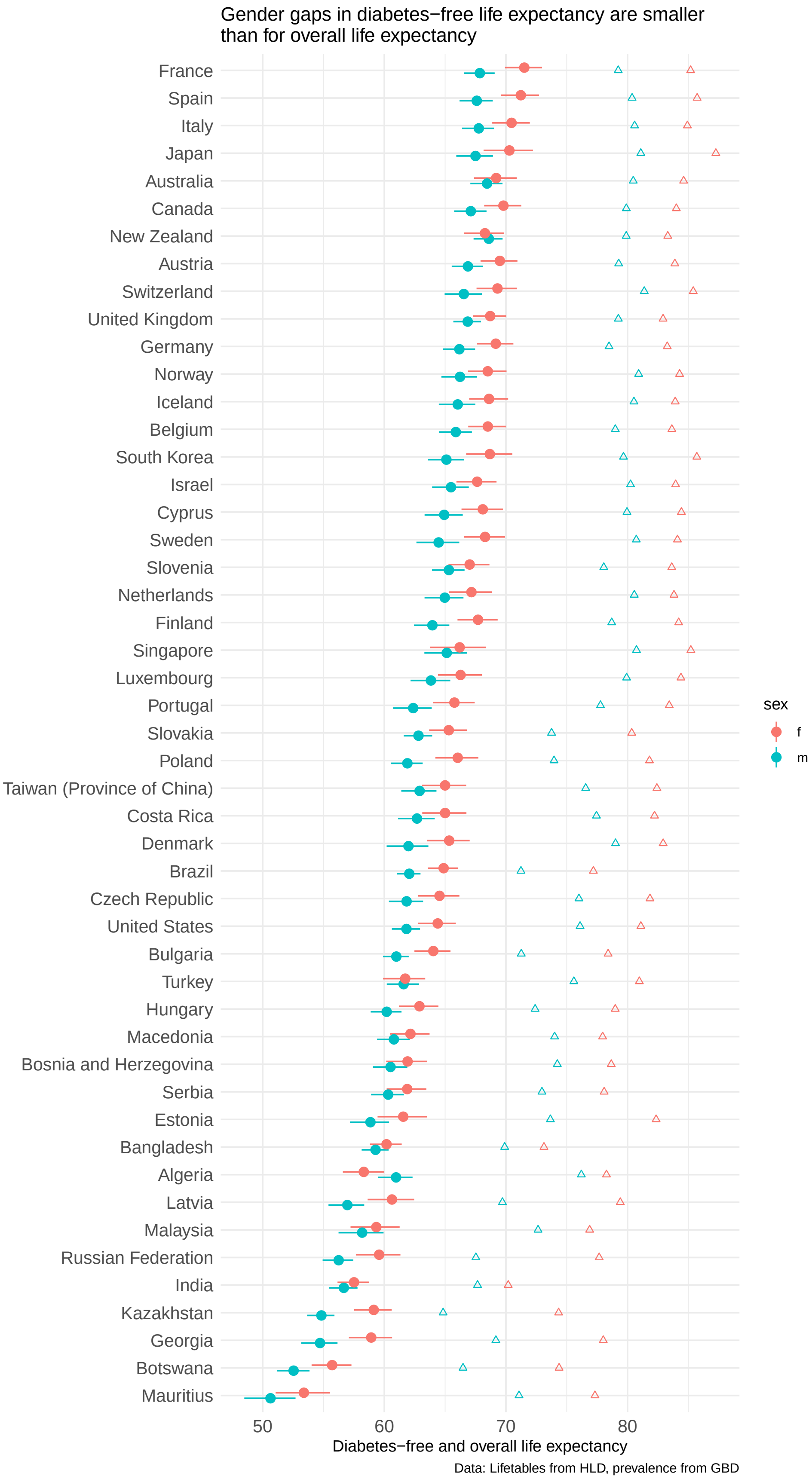
calculate HLE

```
HLExp <-
  HLE %>%
  mutate(Hx = nLx * (1 - val),

         # think this step through by looking at the picture
         Hx_lower = nLx * (1 - upper),
         Hx_upper = nLx * (1 - lower)) %>%
  group_by(location, sex, cause) %>%
  summarize(LE = sum(nLx),
            HLE = sum(Hx),
            HLE_upper = sum(Hx_upper),
            HLE_lower = sum(Hx_lower),
            .groups = "drop")
```

Visualize

```
HLExp %>%
  filter(cause == "Diabetes and kidney diseases",
         !location %in% c("Malta","Tajikistan")) %>%
  ggplot(aes(x = HLE,
             y = reorder(location, HLE),
             color = sex)) +
  geom_point(position = position_dodge2(width = .4, reverse = TRUE)) +
  geom_pointrange(aes(xmin = HLE_lower, xmax = HLE_upper),
                  position = position_dodge2(width = .4, reverse = TRUE)) +
  geom_point(data = filter(HLExp, cause == "Diabetes and kidney diseases",
         !location %in% c("Malta","Tajikistan")),
         mapping = aes(x = LE,
                       y = reorder(location, HLE),
                       color = sex),
         shape = 2) +
  theme_minimal() +
  theme(axis.text = element_text(size = 12)) +
  labs(x = "Diabetes-free and overall life expectancy",
       y = "",
       title = "Gender gaps in diabetes-free life expectancy are smaller
than for overall life expectancy",
caption = "Data: Lifetables from HLD, prevalence from GBD")
```



Gender gaps in diabetes−free life expectancy are smaller than for overall life expectancy

Data: Lifetables from HLD, prevalence from GBD