Cancer Epidemiology

by Dr Jason Wong

Date: 16-11-2022 (Updated 21-11-2022)

The RMarkdown notebook to run your own code can be downloaded here

1. Scenario

You are a grants officer working for the Hong Kong Health Bureau. The HK government has recently announced new special funding in cancer research to be administered by the Bureau. You are tasked with coming up with a proposal for distribution of funding to specific cancer types that are in most need of research.

Discussion points:

- What is important to consider when selecting a cancer type in need of research?
- What type of data is required?

2. Hong Kong population

You are aware that generally cancer is disease that affects the elderly more than the young. You decide to first take a closer look at the structure of the population of Hong Kong.

Historic population of Hong Kong can be obtained from the Census and Statistics Department.

An abridged version of the full historic population of Hong Kong is provided here containing the population of Hong Kong from 1965, 1975, 1985, 1995, 2005, 2015 and 2022 categorised by sex and age (0-19, 20-44, 45-64, 65+).

Discussion points:

- What is the trend in Hong Kong's population over the past ~60 years?
- What is the best way to visualise this data?

Download population data

```
HKPop<- read.table("https://github.com/StatBiomed/BMDS-book/raw/main/notebooks/module5-epidemi/HK_popula
    sep = "\t", header = TRUE, stringsAsFactor=FALSE, check.names = FALSE)
HKPop
```

```
##      Sex   Age  1965  1975   1985   1995   2005   2015   2022
## 1   male  0-19 901.9 989.7  909.0  835.6  719.0  613.8  530.6
## 2   male 20-44 607.4 795.6 1210.2 1363.7 1263.7 1146.9 1039.2
## 3   male 45-64 266.5 414.3  525.9  615.7  896.6 1084.9 1046.2
## 4   male   65+  42.2  84.6  170.5  269.3  384.7  520.0  713.6
## 5 female  0-19 865.6 938.8  840.3  780.7  684.1  575.0  502.1
## 6 female 20-44 539.3 685.6 1093.7 1423.6 1530.4 1531.8 1338.7
## 7 female 45-64 286.3 400.7  470.7  535.0  884.7 1224.3 1314.7
## 8 female   65+  88.7 152.3  235.9  332.5  450.0  594.6  806.5
```

Format data for plotting

Here we convert the original dataframe into simplified format for ggplot2.

```
male<-data.frame(year = as.numeric(colnames(HKPop[1,3:9])),
                 `0-19`  = as.numeric(HKPop[1,3:9]),
                 `20-44` = as.numeric(HKPop[2,3:9]),
                 `45-64` = as.numeric(HKPop[3,3:9]),
                 `65+` = as.numeric(HKPop[4,3:9]),
                 check.names = FALSE)



female<-data.frame(year = as.numeric(colnames(HKPop[1,3:9])),
                 `0-19`  = as.numeric(HKPop[5,3:9]),
                 `20-44` = as.numeric(HKPop[6,3:9]),
                 `45-64` = as.numeric(HKPop[7,3:9]),
                 `65+` = as.numeric(HKPop[8,3:9]),
                 check.names = FALSE)

if (!require("tidyverse")) install.packages("tidyverse")
male<-as_tibble(male) %>% select(year,`0-19`,`20-44`,`45-64`,`65+`) %>% gather (key="age",value="popula
female<-as_tibble(female) %>% select(year,`0-19`,`20-44`,`45-64`,`65+`) %>% gather (key="age",value="po
male
```

```
## # A tibble: 28 x 3
##     year age   `population ('000s)`
##    <dbl> <chr>                <dbl>
##  1  1965 0-19                  902.
##  2  1975 0-19                  990.
##  3  1985 0-19                  909
##  4  1995 0-19                  836.
##  5  2005 0-19                  719
##  6  2015 0-19                  614.
##  7  2022 0-19                  531.
##  8  1965 20-44                 607.
##  9  1975 20-44                 796.
## 10  1985 20-44                1210.
## # ... with 18 more rows
```

```
female
```

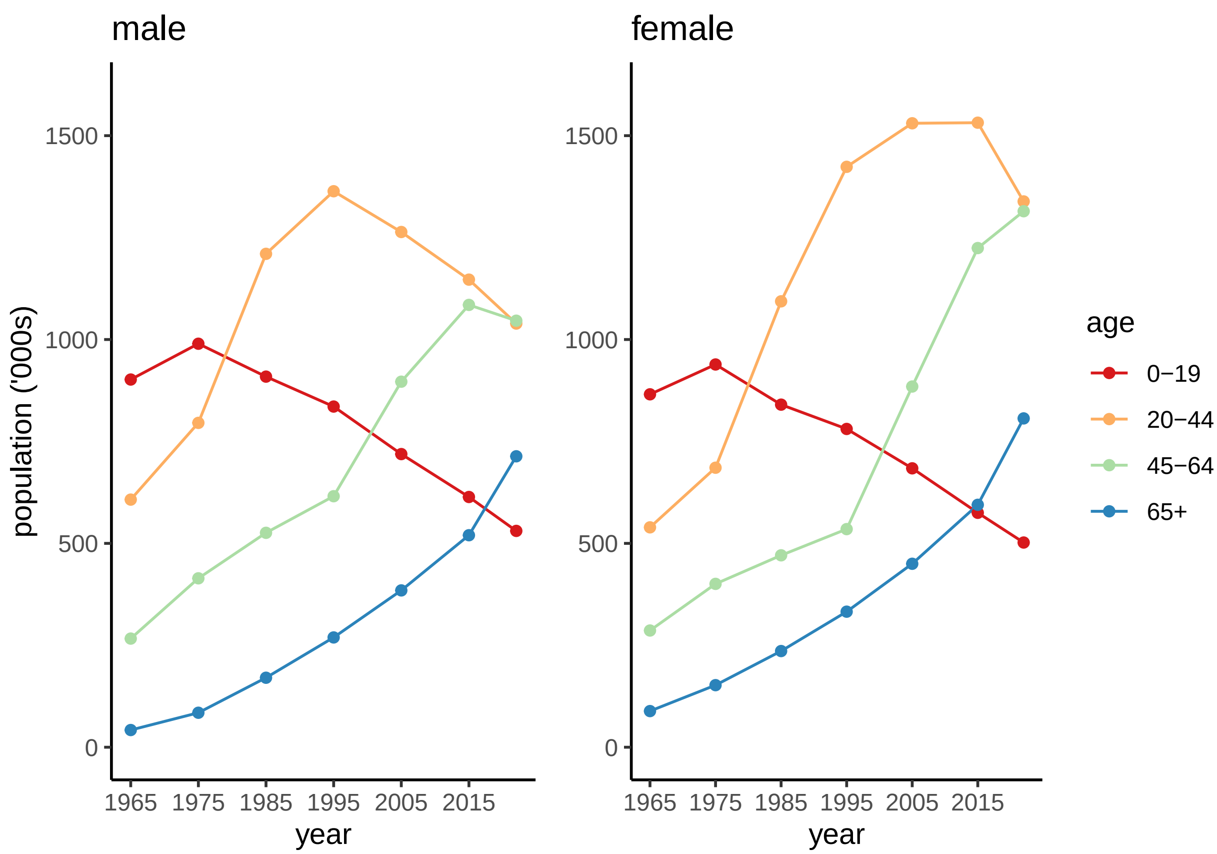
```
## # A tibble: 28 x 3
##     year age   `population ('000s)`
##    <dbl> <chr>                <dbl>
##  1  1965 0-19                  866.
##  2  1975 0-19                  939.
##  3  1985 0-19                  840.
##  4  1995 0-19                  781.
##  5  2005 0-19                  684.
##  6  2015 0-19                  575
##  7  2022 0-19                  502.
##  8  1965 20-44                 539.
```

```
## 9  1975 20-44                686.
## 10  1985 20-44               1094.
## # ... with 18 more rows
```

Plotting population data

Uses ggplot2 and gridExtra to make line plot of male and female population data side-by-side.

```r
if (!require("ggplot2")) install.packages("gglot2")
if (!require("gridExtra")) install.packages("gridExtra")
library(ggplot2)
library(gridExtra)
#Import the necessary packages and libraries
pmale<-ggplot(male,aes(x=year,y=`population ('000s)`,group=age))+
  geom_line(aes(color=age))+
  geom_point(aes(color=age))+
  scale_color_brewer(palette="Spectral")+
  theme_classic()+
  ylim(0,1600)+
  theme(legend.position="none")+
  scale_x_continuous(breaks = seq(1965, 2022, by = 10))+
  ggtitle("male")
pfemale<-ggplot(female,aes(x=year,y=`population ('000s)`,group=age))+
  geom_line(aes(color=age))+
  geom_point(aes(color=age))+
  scale_color_brewer(palette="Spectral")+
  theme_classic()+
  ylim(0,1600)+
  theme(legend.position="right",axis.title.y = element_blank())+
  scale_x_continuous(breaks = seq(1965, 2022, by = 10))+
  ggtitle("female")
grid.arrange(pmale,pfemale,ncol=2,widths=c(3,3.75))
```



3. Cancer registry data

It is clear that Hong Kong has an aging population, thus cancer incidence would also likely increase. To examine cancer incidence and mortality in Hong Kong you obtain data from the Hong Kong Cancer Registry, which is maintained by the Hospital Authority.

Cancer incidence data was summarised for the last three decades (1990-1999, 2000-2009 and 2010-2019).

Discussion points:

- Has incidence been increasing for most cancers? How about mortality?
- How has cancer risk and mortality rate changed in the past 3 decades?
- Has the incidence-to-mortality ratio been decreasing generally? Is it statistically significant?
- Which cancer type has the highest incidence in children (0-19) when compared with the elderly (65+). Is this statistically significantly different to incidence of children versus elderly cancers in general?

Download cancer registry data

```
HKCancer<- read.table("https://github.com/StatBiomed/BMDS-book/raw/main/notebooks/module5-epidemi/HK_ca
                      sep = "\t", header = TRUE, stringsAsFactor=FALSE)
HKCancer
```

```
##         Type    Sex   Age      Year Biliary Bladder Brain Breast Cervix
## 1  incidence   male  0-19 1990-1999       0       6   212      0      0
## 2  incidence   male 20-44 1990-1999      36     165   368      6      0
## 3  incidence   male 45-64 1990-1999     322    1266   393     28      0
## 4  incidence   male   65+ 1990-1999     830    2920   318     35      0
## 5  incidence   male  0-19 2000-2009       0       2   206      0      0
## 6  incidence   male 20-44 2000-2009      31     103   267      9      0
## 7  incidence   male 45-64 2000-2009     297     900   372     52      0
## 8  incidence   male   65+ 2000-2009    1134    3103   299     96      0
## 9  incidence   male  0-19 2010-2019       0       2   155      0      0
## 10 incidence   male 20-44 2010-2019      28      18   276     14      0
## 11 incidence   male 45-64 2010-2019     530     648   511     72      0
## 12 incidence   male   65+ 2010-2019    1568    2399   380    107      0
## 13 incidence female  0-19 1990-1999       0       2   149      6      0
## 14 incidence female 20-44 1990-1999      52      62   283   4228   1263
## 15 incidence female 45-64 1990-1999     246     246   238   5311   1839
## 16 incidence female   65+ 1990-1999     909    1199   284   4122   1574
## 17 incidence female  0-19 2000-2009       0       0   132      4      4
## 18 incidence female 20-44 2000-2009      34      39   223   5887   1204
## 19 incidence female 45-64 2000-2009     287     179   265  11833   1676
## 20 incidence female   65+ 2000-2009    1190    1090   237   5774   1329
## 21 incidence female  0-19 2010-2019       0       2   119      1      0
## 22 incidence female 20-44 2010-2019      36      17   223   6658   1294
## 23 incidence female 45-64 2010-2019     452     188   394  22096   2288
## 24 incidence female   65+ 2010-2019    1500     872   283  10337   1269
## 25 mortality   male  0-19 1990-1999       0       0    76      0      0
## 26 mortality   male 20-44 1990-1999      14       6   156      0      0
## 27 mortality   male 45-64 1990-1999     169     240   268      5      0
## 28 mortality   male   65+ 1990-1999     548    1036   240     13      0
## 29 mortality   male  0-19 2000-2009       0       0    63      0      0
## 30 mortality   male 20-44 2000-2009      10       9    99      1      0
## 31 mortality   male 45-64 2000-2009     111     175   205      3      0
## 32 mortality   male   65+ 2000-2009     592    1209   209     26      0
## 33 mortality   male  0-19 2010-2019       1       0    28      0      0
## 34 mortality   male 20-44 2010-2019       5       4   107      2      0
## 35 mortality   male 45-64 2010-2019     181     166   288      8      0
## 36 mortality   male   65+ 2010-2019     595    1272   206     25      0
## 37 mortality female  0-19 1990-1999       0       0    59      0      0
## 38 mortality female 20-44 1990-1999      20       6   104    644    193
## 39 mortality female 45-64 1990-1999     141      40   113   1284    494
## 40 mortality female   65+ 1990-1999     663     486   197   1504    748
## 41 mortality female  0-19 2000-2009       0       0    42      0      0
## 42 mortality female 20-44 2000-2009      12       3    77    691    162
## 43 mortality female 45-64 2000-2009     125      31   145   2112    415
## 44 mortality female   65+ 2000-2009     731     482   158   1809    670
## 45 mortality female  0-19 2010-2019       0       0    29      0      0
## 46 mortality female 20-44 2010-2019       7       2    67    525    167
## 47 mortality female 45-64 2010-2019     170      51   196   3452    622
## 48 mortality female   65+ 2010-2019     782     534   174   2600    709
##    Colorectum Eye Hodgkin.lymphoma Kaposi Kidney Larynx Leukaemia Liver  Lung
## 1          15  29               35      0     24      1       356    58     6
## 2        1087   8               64      0    141     86       509  1741  1036
## 3        4644   9               47      0    530    824       453  5702  8420
## 4        8058  21               41      0    721   1265       623  4806 15152
```

```
## 5          5  38           32     0    20     0       329    33     3
## 6        957   8          119     9   221    37       437  1116   772
## 7       6418  18           77     4  1071   658       642  6015  7909
## 8      13352  13           77     8  1275  1125       864  5808 18870
## 9          6  26           53     0    31     0       338    24     2
## 10      1005  10          163    25   293    29       401   659   541
## 11     10034  28           87    18  1978   684      1033  6500  9523
## 12     18003  15          114    12  1977  1043      1381  6716 20588
## 13         2  33           10     0    25     0       296    37     7
## 14       932  11           61     0    98    13       393   353   597
## 15      3166   9           19     0   256    59       312  1093  2629
## 16      7728  16           33     0   503   130       534  2244  8428
## 17         5  14           30     0    15     0       219    13     5
## 18       938  12          122     2   120     5       405   212   625
## 19      4430  10           33     0   482    37       469  1041  3397
## 20     10810   8           35     0   830    85       723  2637  9517
## 21         5  14           32     0    19     0       237    21     0
## 22      1057   9          174     0   135     7       443   168   647
## 23      7200  16           58     4   887    47       786  1168  6312
## 24     13119  20           58     5  1203    84       913  3007 10872
## 25        10   3            0     0     2     0       120    36     1
## 26       347   1           11     0    47    16       297  1059   605
## 27      1708   2           12     0   208   248       296  4118  5828
## 28      3969   3           16     0   385   575       471  4161 12819
## 29         4   2            0     0     2     0        81    19     0
## 30       289   1            9     1    34     7       209   777   514
## 31      2116   5           11     0   312   160       378  4432  5894
## 32      6541   3           23     1   628   501       766  5658 17272
## 33         1   3            0     0     3     0        62    11     0
## 34       221   2            2     4    36     6       152   405   308
## 35      2950   8           11     2   447   151       491  4291  6527
## 36      8484   2           35     0   895   426      1138  6478 18695
## 37         1   3            0     0     4     0        93    15     2
## 38       302   1            3     0    22     0       215   187   379
## 39      1100   1            1     0    81    14       212   737  1821
## 40      3633   3           11     0   265    49       452  1915  7182
## 41         1   2            1     0     1     0        46     4     2
## 42       266   6            5     0    23     1       165   143   383
## 43      1341   5            4     0   127     9       288   717  2163
## 44      5204   1           14     0   467    46       655  2712  8732
## 45         4   0            1     0     2     0        42     6     2
## 46       218   0            5     0    14     0       145    92   304
## 47      1955   0            7     1   166     2       311   759  3460
## 48      6640   0           15     0   564    32       781  3315  9399
##    Melanoma Mesothelioma Multiple.myeloma Nasal Nasopharynx
## 1         3            0                1     4          33
## 2        50            0               43    53        3049
## 3        82            0              210   109        3724
## 4       107            0              421   120        1212
## 5         1            0                1     7          16
## 6        41            5               38    38        2114
## 7        98           34              326   161        3750
## 8       137           77              674   137        1177
## 9         0            0                0     3          21
```

```
## 10       47            4               26    41        1339
## 11      186           51              525   174        3653
## 12      195          154              910   154        1207
## 13        2            0                1     1          13
## 14       44            0               19    31        1380
## 15       58            0              150    57        1157
## 16      102            0              417    93         536
## 17        2            0                0     5           6
## 18       41            8               21    36         942
## 19      102           12              185    74        1187
## 20      118           14              559    92         465
## 21        3            0                0     6           6
## 22       70           10               24    37         583
## 23      128           34              399   105        1145
## 24      179           24              636   102         389
## 25        2            0                0     0           6
## 26       17            0               18    11         771
## 27       46            0              122    42        1673
## 28       72            0              297    55         735
## 29        0            0                0     1           2
## 30       16            5               17     9         452
## 31       55           24              137    32        1551
## 32       85           54              493    62         867
## 33        0            0                0     0           2
## 34       12            0                5     2         239
## 35       79           38              190    50        1413
## 36      128          140              571    60         812
## 37        0            0                0     0           1
## 38       14            0                7     7         242
## 39       25            0               74    12         421
## 40       50            0              273    49         302
## 41        1            0                0     0           0
## 42       11            3                6     5         159
## 43       49            9               88    14         385
## 44       89           16              388    40         305
## 45        1            0                0     0           1
## 46       23            7                4     1          77
## 47       73           13              101    17         345
## 48      120           23              456    36         268
##     Non.Hodgkin.lymphoma Non.melanoma.skin Oesophagus Oral Ovary Pancreas Penis
## 1                    137                 8          0   18     0        4     2
## 2                    596               211        158  324     0      106    19
## 3                    997               613       1873 1165     0      514    78
## 4                   1210               968       2203 1055     0      887   184
## 5                    111                 4          0    6     0        1     0
## 6                    454               220         82  312     0       71    18
## 7                   1216               917       1528 1354     0      780    71
## 8                   1756              1835       2126 1416     0     1440   199
## 9                    106                 1          0    9     0        1     0
## 10                   487               291         54  304     0      109    21
## 11                  1910              1782       1321 1941     0     1402   154
## 12                  2633              2975       1982 1813     0     2302   279
## 13                    75                 2          0   21    77        2     0
## 14                   444               156         48  241   854       60     0
```

```
## 15                  594               374        254  364   970      290     0
## 16                 1120              1249        761  485   776      875     0
## 17                   43                 4          1   20   108        0     0
## 18                  461               182         13  254  1299       62     0
## 19                  888               580        190  454  1846      461     0
## 20                 1409              2476        725  768   891     1329     0
## 21                   57                 4          0    9    84        3     0
## 22                  475               303         16  259  1466       92     0
## 23                 1561              1168        174  920  3166      923     0
## 24                 1883              3268        624 1091  1128     2101     0
## 25                   25                 0          0    1     0        3     0
## 26                  179                 4         86   79     0       78     3
## 27                  380                18       1341  501     0      476    10
## 28                  651                47       1652  545     0      918    31
## 29                   17                 0          0    1     0        0     1
## 30                  145                 5         53   62     0       49     3
## 31                  461                21       1010  465     0      680     9
## 32                 1119                72       1798  743     0     1339    43
## 33                   15                 0          0    1     0        0     0
## 34                   83                 2         30   63     0       60     1
## 35                  542                25        914  563     0     1100    18
## 36                 1473                93       1640  824     0     2191    39
## 37                   16                 1          0    1     1        1     0
## 38                  118                 2         12   30   139       33     0
## 39                  194                 6        118   75   431      266     0
## 40                  605                42        505  193   520      875     0
## 41                    8                 2          0    1     3        0     0
## 42                  104                 3          9   40   155       42     0
## 43                  256                 4        103  100   601      382     0
## 44                  895                72        585  297   649     1317     0
## 45                    5                 0          0    2     1        0     0
## 46                   81                 1          5   32   167       50     0
## 47                  312                16         93  170  1056      713     0
## 48                 1133                72        518  400   879     2075     0
##    Placenta Prostate Sarcoma Small.intestine Stomach Testis Thymus Thyroid
## 1         0        0     137               0       5     61     12      22
## 2         0       10     320              31     429    275     71     250
## 3         0      481     281             102    1990     80     70     288
## 4         0     3073     261             131    3634    129     59     194
## 5         0        0     118               0       6     57     16      21
## 6         0        6     307              34     331    395     96     333
## 7         0     1588     367             113    1954     65    137     447
## 8         0     8676     316             163    4256     44     94     278
## 9         0        1     110               1       2     53     23      17
## 10        0       18     299              46     192    587     68     496
## 11        0     4128     527             293    2158    100    137     834
## 12        0    14703     507             336    4720     15     85     424
## 13        1        0     128               1       1      0      3      80
## 14       13        0     283              22     492      0     38    1294
## 15        0        0     232              53     826      0     38     788
## 16        0        0     232             116    2319      0     38     479
## 17        0        0     105               0       2      0      1      81
## 18        8        0     322              24     379      0     64    1659
## 19        3        0     300              80     971      0    116    1570
```

```
## 20          0          0        247                   153      2512        0        65      595
## 21          0          0         86                     1         0        0         5      103
## 22         12          0        330                    38       318        0        30     2291
## 23          3          0        509                   221      1653        0        81     3200
## 24          0          0        351                   251      2809        0        67      940
## 25          0          0         41                     1         0        2         7        1
## 26          0          3         98                     8       220       23        28        4
## 27          0        131         99                    40      1074        5        38       34
## 28          0       1158        153                    53      2452       15        49       70
## 29          0          0         30                     0         1        1         2        0
## 30          0          2         94                     6       140        9        21       10
## 31          0        160        116                    34      1046        5        42       35
## 32          0       2299        161                    89      2814        8        58      101
## 33          0          0         27                     0         0        1         2        1
## 34          0          3         89                     4        97       11        20        6
## 35          0        279        209                    57       996        4        81       45
## 36          0       3638        336                   151      2955        2        74      151
## 37          0          0         39                     0         1        0         2        0
## 38          7          0         68                     6       269        0         9       10
## 39          0          0         97                    15       451        0        11       47
## 40          0          0        149                    52      1544        0        38      175
## 41          0          0         29                     0         0        0         0        0
## 42          4          0         65                     6       210        0        13        8
## 43          1          0        106                    26       561        0        32       40
## 44          0          0        152                    95      1712        0        44      215
## 45          0          0         20                     0         0        0         0        1
## 46          4          0         82                     1       158        0        12        6
## 47          3          0        219                    35       731        0        43       44
## 48          0          0        236                   109      1818        0        50      211
##    Uterus Vulva
## 1       0     0
## 2       0     0
## 3       0     0
## 4       0     0
## 5       0     0
## 6       0     0
## 7       0     0
## 8       0     0
## 9       0     0
## 10      0     0
## 11      0     0
## 12      0     0
## 13      2     1
## 14    478    49
## 15   1362    96
## 16    729   235
## 17      2     2
## 18    821    44
## 19   3071   147
## 20   1130   336
## 21      3     1
## 22   1165    57
## 23   6582   235
## 24   1864   486
```

```
## 25      0      0
## 26      0      0
## 27      0      0
## 28      0      0
## 29      0      0
## 30      0      0
## 31      0      0
## 32      0      0
## 33      0      0
## 34      0      0
## 35      0      0
## 36      0      0
## 37      0      0
## 38     18      4
## 39    119     22
## 40    149     71
## 41      0      0
## 42     23     16
## 43    198     53
## 44    205    133
## 45      0      0
## 46     43     18
## 47    517    107
## 48    432    197
```

3a. Visualise changes in incidence and mortality

```
# first plot incidence for each cancer type

#Import the necessary packages and libraries
if (!require("ggplot2")) install.packages("gglot2")
if (!require("gridExtra")) install.packages("gridExtra")
if (!require("ggpubr")) install.packages("ggpubr")
library(ggplot2)
library(gridExtra)
library(ggpubr)

HKCancer_inc <- HKCancer[HKCancer$Type=='incidence',]

p<-list()

for (i in 1:(ncol(HKCancer_inc)-4)){
  p[[i]]<-ggplot(HKCancer_inc,aes_string(fill=names(HKCancer_inc)[3],x=names(HKCancer_inc)[4],
                                         y=names(HKCancer_inc)[i+4],group=names(HKCancer_inc)[3]))+
       geom_bar(position="dodge",stat="identity")+
       facet_wrap(~Sex) +
       theme_classic()+
       scale_fill_brewer(palette="Spectral")+
       theme(legend.position="none")+
       theme(text = element_text(size = 10))+
       ggtitle(names(HKCancer_inc)[i+4])+
       ylab("Incidence")
```

```
}
do.call('grid.arrange',c(p,ncol=3,nrow=12))
```


Biliary
Bladder
Brain
Breast
Cervix
Colorectum
Eye
Hodgkin.lymphoma
Kaposi
Kidney
Larynx
Leukaemia
Liver
Lung
Melanoma
Mesothelioma
Multiple.myeloma
Nasal
Nasopharynx
Non.Hodgkin.lymphoma
Non.melanoma.skin
Oesophagus
Oral
Ovary
Pancreas
Penis
Placenta
Prostate
Sarcoma
Small.intestine
Stomach
Testis
Thymus
Thyroid
Uterus
Vulva
female
male
Incidence
Year
1990−1999
2000−2009
2010−2019

```r
# now plot mortality for each cancer type

#Import the necessary packages and libraries
if (!require("ggplot2")) install.packages("gglot2")
if (!require("gridExtra")) install.packages("gridExtra")
if (!require("ggpubr")) install.packages("ggpubr")
if (!require("dplyr")) install.packages("dplyr")
library(ggplot2)
library(gridExtra)
library(ggpubr)
library(dplyr)

HKCancer_mort <- HKCancer[HKCancer$Type=='mortality',]

p<-list()
#(ncol(HKCancer_inc)-4)
for (i in 1:(ncol(HKCancer_mort)-4)){
  p[[i]]<-ggplot(HKCancer_mort,aes_string(fill=names(HKCancer_mort)[3],x=names(HKCancer_mort)[4],
                                           y=names(HKCancer_mort)[i+4],group=names(HKCancer_mort)[3]))+
       geom_bar(position="dodge",stat="identity")+
       facet_wrap(~Sex) +
       theme_classic()+
       scale_fill_brewer(palette="Spectral")+
       theme(legend.position="none")+
       theme(text = element_text(size = 10))+
       ggtitle(names(HKCancer_mort)[i+4])+
       ylab("Mortality")
}
do.call('grid.arrange',c(p,ncol=3,nrow=12))
```


Biliary
Bladder
Brain
Breast
Cervix
Colorectum
Eye
Hodgkin.lymphoma
Kaposi
Kidney
Larynx
Leukaemia
Liver
Lung
Melanoma
Mesothelioma
Multiple.myeloma
Nasal
Nasopharynx
Non.Hodgkin.lymphoma
Non.melanoma.skin
Oesophagus
Oral
Ovary
Pancreas
Penis
Placenta
Prostate
Sarcoma
Small.intestine
Stomach
Testis
Thymus
Thyroid
Uterus
Vulva
female
male
Mortality
Year
1990−1999
2000−2009
2010−2019

3b. Calculate cancer risk and mortality rate

To calculate disease risk we need to calculated the number of new cases over the number of persons at risk over a specific time period. We have the incidence for each decade and can estimate the number of persons at risk based on the population in 1995, 2005 and 2015.

```
# cancer risk calculation

HKCancer_inc <- HKCancer[HKCancer$Type=='incidence',]
HKCancer_inc_risk <- HKCancer_inc[,1:5]

risk <- function(x,age,sex,year){
  if (is.integer(x)){
    pop_n <- HKPop %>% filter(Sex==sex & Age==age)
    if (year == "1990-1999"){ return (as.double(x)/as.double(pop_n$`1995`)) }
    else if (year == "2000-2009") {return (as.double(x)/as.double(pop_n$`2005`))}
    else { return (as.double(x)/as.double(pop_n$`2015`)) }
  }
  return (x)
}

for (i in 5:ncol(HKCancer_inc)){
  HKCancer_inc_risk[names(HKCancer_inc)[i]] <- mapply(risk,HKCancer_inc[,i],HKCancer_inc[,3],HKCancer_in
}
```

```
#visualise cancer risk

p<-list()
for (i in 1:(ncol(HKCancer_inc_risk)-4)){
  p[[i]]<-ggplot(HKCancer_inc_risk,aes_string(fill=names(HKCancer_inc_risk)[3],x=names(HKCancer_inc_risk
                                              y=names(HKCancer_inc_risk)[i+4],group=names(HKCancer_inc_r
       geom_bar(position="dodge",stat="identity")+
       facet_wrap(~Sex) +
       theme_classic()+
       scale_fill_brewer(palette="Spectral")+
       theme(legend.position="none")+
       theme(text = element_text(size = 10))+
       ggtitle(names(HKCancer_inc)[i+4])+
       ylab("incidence per 1000")
}
do.call('grid.arrange',c(p,ncol=3,nrow=12))
```


Biliary
Bladder
Brain
Breast
Cervix
Colorectum
Eye
Hodgkin.lymphoma
Kaposi
Kidney
Larynx
Leukaemia
Liver
Lung
Melanoma
Mesothelioma
Multiple.myeloma
Nasal
Nasopharynx
Non.Hodgkin.lymphoma
Non.melanoma.skin
Oesophagus
Oral
Ovary
Pancreas
Penis
Placenta
Prostate
Sarcoma
Small.intestine
Stomach
Testis
Thymus
Thyroid
Uterus
Vulva
female
male
incidence per 1000
Year
1990–1999
2000–2009
2010–2019

```r
# mortality rate calculation

HKCancer_mort <- HKCancer[HKCancer$Type=='mortality',]
HKCancer_mort_risk <- HKCancer_mort[,1:5]

risk <- function(x,age,sex,year){
  if (is.integer(x)){
    pop_n <- HKPop %>% filter(Sex==sex & Age==age)
    if (year == "1990-1999"){ return (as.double(x)/as.double(pop_n$`1995`)) }
    else if (year == "2000-2009") {return (as.double(x)/as.double(pop_n$`2005`))}
    else { return (as.double(x)/as.double(pop_n$`2015`)) }
  }
  return (x)
}

for (i in 5:ncol(HKCancer_mort)){
  HKCancer_mort_risk[names(HKCancer_mort)[i]] <- mapply(risk,HKCancer_mort[,i],HKCancer_mort[,3],HKCanc
}
```

```r
#visualise mortality rate

p<-list()
for (i in 1:(ncol(HKCancer_mort_risk)-4)){
  p[[i]]<-ggplot(HKCancer_mort_risk,aes_string(fill=names(HKCancer_mort_risk)[3],x=names(HKCancer_mort_r
                                               y=names(HKCancer_mort_risk)[i+4],group=names(HKCancer_mor
       geom_bar(position="dodge",stat="identity")+
       facet_wrap(~Sex) +
       theme_classic()+
       scale_fill_brewer(palette="Spectral")+
       theme(legend.position="none")+
       theme(text = element_text(size = 10))+
       ggtitle(names(HKCancer_inc)[i+4])+
       ylab("mortality per 1000")
}
do.call('grid.arrange',c(p,ncol=3,nrow=12))
```


Biliary
Bladder
Brain
Breast
Cervix
Colorectum
Eye
Hodgkin.lymphoma
Kaposi
Kidney
Larynx
Leukaemia
Liver
Lung
Melanoma
Mesothelioma
Multiple.myeloma
Nasal
Nasopharynx
Non.Hodgkin.lymphoma
Non.melanoma.skin
Oesophagus
Oral
Ovary
Pancreas
Penis
Placenta
Prostate
Sarcoma
Small.intestine
Stomach
Testis
Thymus
Thyroid
Uterus
Vulva
female
male
mortality per 1000
1990–1999
2000–2009
2010–2019
Year

3c. Mortality-incidence ratio

Cancer research can be focused on improving cancer outcomes in a number of ways. For example cancer prevention research that seeks to reduce cancer incidence which would also ultimately reduce cancer mortality. Another area is cancer therapy which would not affect incidence but seeks to reduce mortality, or at least prolong survival. We don't go into survival analysis in this tutorial, but a way to get an idea whether treatment is improving by looking at the mortality-incidence ratio.

```
HKCancer_inc <- HKCancer[HKCancer$Type=='incidence',]
HKCancer_mort <- HKCancer[HKCancer$Type=='mortality',]
HKCancer_mort_inc <- HKCancer_mort[,1:5]

risk <- function(mort,inc){
  if (inc == 0){ return (0) }
  return (as.double(mort)/as.double(inc))
}

for (i in 5:ncol(HKCancer_mort)){
  HKCancer_mort_inc[names(HKCancer_mort)[i]] <- mapply(risk,HKCancer_mort[,i],HKCancer_inc[,i])
}
```

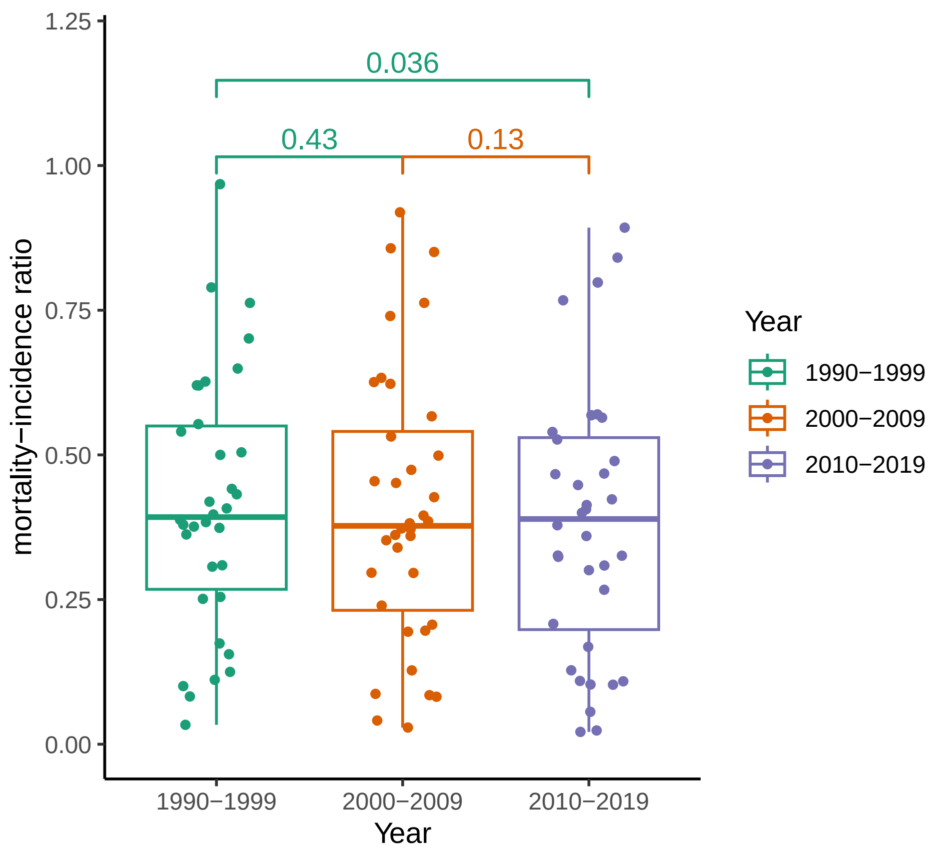
```
#visualise mortality incidence ratio

p<-list()
for (i in 1:(ncol(HKCancer_mort_inc)-4)){
  p[[i]]<-ggplot(HKCancer_mort_inc,aes_string(fill=names(HKCancer_mort_inc)[3],x=names(HKCancer_mort_in
                                              y=names(HKCancer_mort_inc)[i+4],group=names(HKCancer_mort_
       geom_bar(position="dodge",stat="identity")+
       facet_wrap(~Sex) +
       theme_classic()+
       scale_fill_brewer(palette="Spectral")+
       theme(legend.position="none")+
       theme(text = element_text(size = 10))+
       ggtitle(names(HKCancer_inc)[i+4])+
       ylab("mortality-incidence ratio")
}
do.call('grid.arrange',c(p,ncol=3,nrow=12))
```


Biliary
Bladder
Brain
Breast
Cervix
Colorectum
Eye
Hodgkin.lymphoma
Kaposi
Kidney
Larynx
Leukaemia
Liver
Lung
Melanoma
Mesothelioma
Multiple.myeloma
Nasal
Nasopharynx
Non.Hodgkin.lymphoma
Non.melanoma.skin
Oesophagus
Oral
Ovary
Pancreas
Penis
Placenta
Prostate
Sarcoma
Small.intestine
Stomach
Testis
Thymus
Thyroid
Uterus
Vulva
female
male
mortality−incidence ratio
1990−1999
2000−2009
2010−2019
Year

3d. Paired t-test on mortality-incidence ratio change

In general, across the different cancer types is cancer treatment improving? We can use a paired t-test comparing the mortality-incidence ratio of cancers from the 1990-1999 period with the 2010-2019 period.

```
#First sum up all incidence and mortality data for each cancer type across age and sex

HKCancer_inc_sum <- aggregate(HKCancer_inc[,-(1:4)],list(HKCancer_inc$Year),FUN=sum)
HKCancer_mort_sum <- aggregate(HKCancer_mort[,-(1:4)],list(HKCancer_mort$Year),FUN=sum)
HKCancer_mort_inc_year <- data.frame(Year=HKCancer_mort_sum[,1])

risk <- function(mort,inc){
  return (as.double(mort)/as.double(inc))
}

for (i in 2:ncol(HKCancer_mort_sum)){
  HKCancer_mort_inc_year[names(HKCancer_mort_sum)[i]] <- mapply(risk,HKCancer_mort_sum[,i],HKCancer_inc_
}

HKCancer_mort_inc_year_m <-data.frame(`1990-1999` = as.numeric(HKCancer_mort_inc_year[1,2:37]),
                                      `2000-2009` = as.numeric(HKCancer_mort_inc_year[2,2:37]),
                                      `2010-2019` = as.numeric(HKCancer_mort_inc_year[3,2:37]),
                                      check.names = FALSE)

HKCancer_mort_inc_year_m$Cancer <- names(HKCancer_mort_inc_year)[-1]

HKCancer_mort_inc_year_t<-as_tibble(HKCancer_mort_inc_year_m) %>% select(`Cancer`,`1990-1999`,`2000-200

plot<-ggplot(HKCancer_mort_inc_year_t,aes(x=Year,y=MIR, color=Year))+
       geom_boxplot(na.rm=T) +
       theme_classic()+
       scale_color_brewer(palette="Dark2")+
       geom_jitter(shape=16, position=position_jitter(0.2),na.rm=T)+
       ylab("mortality-incidence ratio")+
       ylim(0,1.2) + geom_signif(comparisons = list(c("1990-1999", "2010-2019")),
                          map_signif_level=F, test= "t.test",test.args = list(paired = TRUE), na
       geom_signif(comparisons = list(c("1990-1999", "2000-2009")), map_signif_level=F, test= "t.test",
       geom_signif(comparisons = list(c("2000-2009", "2010-2019")), map_signif_level=F, test= "t.test",
plot
```




```
p<-list()
p[[1]]<-t.test(HKCancer_mort_inc_year_m$`1990-1999`,HKCancer_mort_inc_year_m$`2000-2009`,paired=TRUE,al
p[[2]]<-t.test(HKCancer_mort_inc_year_m$`1990-1999`,HKCancer_mort_inc_year_m$`2010-2019`,paired=TRUE,al
p[[3]]<-t.test(HKCancer_mort_inc_year_m$`2000-2009`,HKCancer_mort_inc_year_m$`2010-2019`,paired=TRUE,al
p
```

```
## [[1]]
##
##  Paired t-test
##
## data:  HKCancer_mort_inc_year_m$‘1990-1999‘ and HKCancer_mort_inc_year_m$‘2000-2009‘
## t = 0.79695, df = 33, p-value = 0.4312
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.01415536  0.03238636
## sample estimates:
## mean difference
##     0.009115498
##
##
## [[2]]
##
##  Paired t-test
##
## data:  HKCancer_mort_inc_year_m$‘1990-1999‘ and HKCancer_mort_inc_year_m$‘2010-2019‘
```

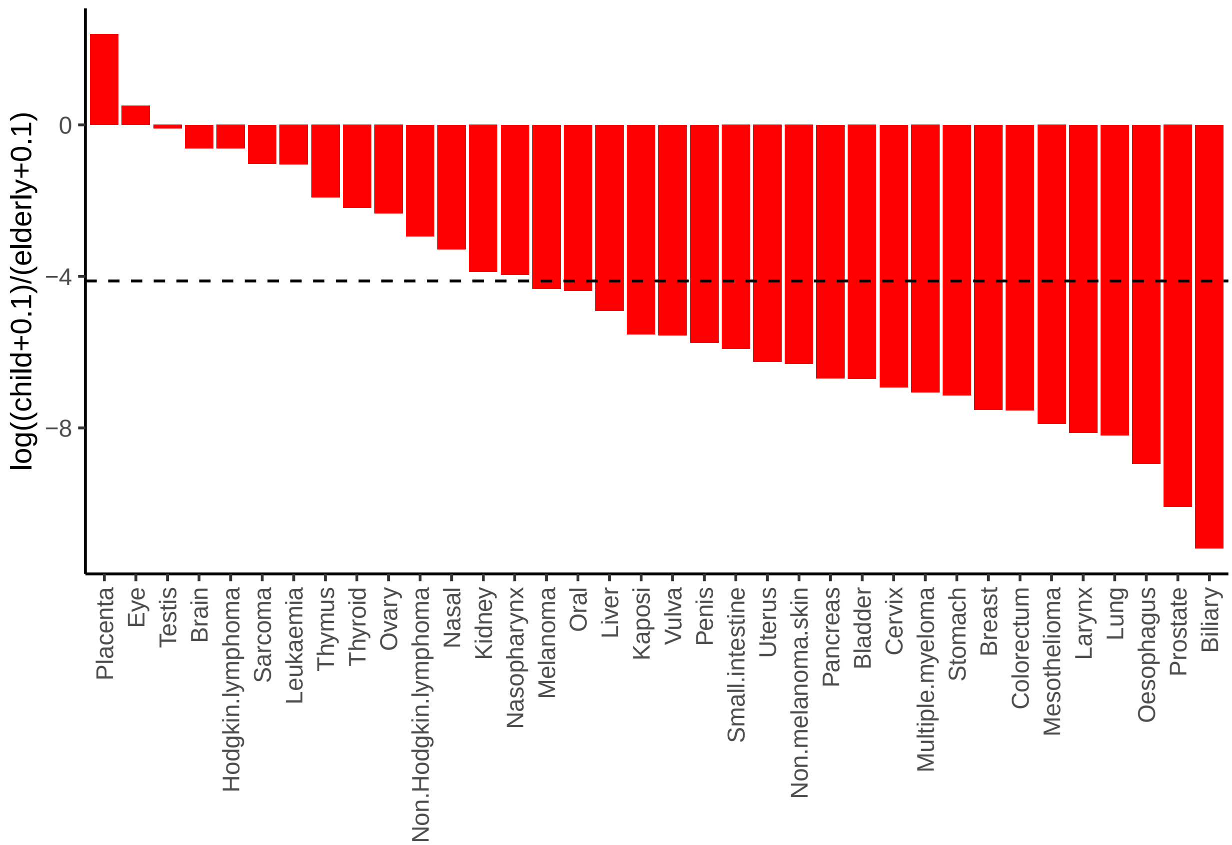
```
## t = 2.1861, df = 33, p-value = 0.036
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.002055217 0.057209924
## sample estimates:
## mean difference
##      0.02963257
##
##
## [[3]]
##
##  Paired t-test
##
## data:  HKCancer_mort_inc_year_m$`2000-2009` and HKCancer_mort_inc_year_m$`2010-2019`
## t = 1.5413, df = 35, p-value = 0.1322
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.005438019  0.039734021
## sample estimates:
## mean difference
##        0.017148
```

3e. Childhood versus elderly cancers

Although it is clear that the incidence of cancer is typically higher in the elderly, some cancers affect children as well. What cancer types disproportionate affect children? For each cancer type, compare the proportion of 0-19 versus 65+ incidence against the 0-19 versus 65+ incidence for all other cancer types.

```
# Examine the proportion of childhood

HKCancer_inc <- HKCancer[HKCancer$Type=='incidence',]
HKCancer_inc_age_sum <- aggregate(HKCancer_inc[,-(1:4)],list(HKCancer_inc$Age),FUN=sum)
HKCancer_inc_age_sum$Total<- rowSums(HKCancer_inc_age_sum[,-1])
HKCancer_inc_age_sum_csq <-  data.frame(cancer=names(HKCancer_inc_age_sum[,2:ncol(HKCancer_inc_age_sum)]

pval <- list()
ratio <- list()
for (i in 2:(ncol(HKCancer_inc_age_sum))){
  val = Map('-',HKCancer_inc_age_sum$Total,HKCancer_inc_age_sum[,i])
  dat <- data.frame(cancer=HKCancer_inc_age_sum[c(1,4),i], other =c(val[[1]],val[[2]]))
  pval <- append(pval,fisher.test(dat)$p.val)
  ratio<- append(ratio,log(as.double(dat[1,1]+0.1)/as.double(dat[2,1]+0.1)))
}
HKCancer_inc_age_sum_csq$pval <- pval
HKCancer_inc_age_sum_csq$ratio <- ratio

plot_child<-ggplot(HKCancer_inc_age_sum_csq[-37,],aes(x=reorder(cancer,-as.numeric(ratio)),y=as.numeric
       geom_bar(stat="identity",fill="red")+
       geom_hline(yintercept=-4.12132318942113, linetype="dashed",
                  color = "black", linewidth=0.5)+
       theme_classic()+
       scale_fill_brewer(palette="Spectral")+
       theme(legend.position="none")+
```

```
       #theme(text = element_text(size = 10))+
       theme(axis.title.x=element_blank())+
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
       ylab("log((child+0.1)/(elderly+0.1)")
plot_child
```




```
HKCancer_inc_age_sum_csq
```

```
##                  cancer          pval       ratio
## 1               Biliary  1.21618e-216   -11.17481
## 2               Bladder 9.601084e-314   -6.711128
## 3                 Brain             0   -0.615666
## 4                Breast             0   -7.519824
## 5                Cervix 1.635299e-126   -6.925188
## 6             Colorectum             0   -7.531208
## 7                   Eye 2.085337e-112   0.5039276
## 8      Hodgkin.lymphoma  3.658856e-81  -0.6227962
## 9                Kaposi       0.41662   -5.525453
## 10               Kidney  1.560619e-71   -3.882371
## 11               Larynx 1.963055e-113   -8.129416
## 12            Leukaemia             0   -1.043172
## 13                Liver             0   -4.909033
## 14                 Lung             0   -8.191896
## 15             Melanoma  6.246412e-15   -4.324192
## 16         Mesothelioma  5.396388e-09   -7.897668
```

```
## 17     Multiple.myeloma 6.126613e-106   -7.062026
## 18                Nasal  0.0001285609   -3.286427
## 19          Nasopharynx  4.628193e-74    -3.95948
## 20 Non.Hodgkin.lymphoma  4.513518e-11   -2.940272
## 21    Non.melanoma.skin             0   -6.315107
## 22           Oesophagus 1.377432e-251   -8.943186
## 23                 Oral 5.100139e-108   -4.379029
## 24                Ovary  0.0002847234    -2.34054
## 25             Pancreas  3.11451e-246   -6.690686
## 26                Penis  3.023792e-18   -5.753479
## 27             Placenta    0.07013259    2.397895
## 28             Prostate             0   -10.08778
## 29              Sarcoma 9.774701e-213   -1.028899
## 30      Small.intestine  1.479593e-31   -5.916202
## 31              Stomach             0   -7.137096
## 32               Testis  1.871184e-97 -0.09472556
## 33               Thymus  6.748934e-06   -1.915502
## 34              Thyroid  7.821825e-09   -2.194891
## 35               Uterus 1.457974e-105   -6.262217
## 36                Vulva  1.521831e-27   -5.552298
## 37                Total             1   -4.121323
```

4. Existing cancer funding and publication data

The Hong Kong government established in Health and Medical Research Fund (HMRF) in 2011 to specifically provide research funding for health and medical research in Hong Kong. Since 2016 over 370 projects in the category of Cancer has been funded for a total of ~$400 M dollars. A list of all funded projects can be found on the Health Bureau webpage. You would like to use this data to see if there is any association between previous project funding and the epidemiology of cancers in Hong Kong.

We can also do a similar thing with publications and ask if the research publications in Hong Kong have been aligned with the incidence and mortality. We can obtain this data from PubMed using the following terms: ("Hong Kong"[Affiliation]) AND (neoplasms[MeSH Terms])

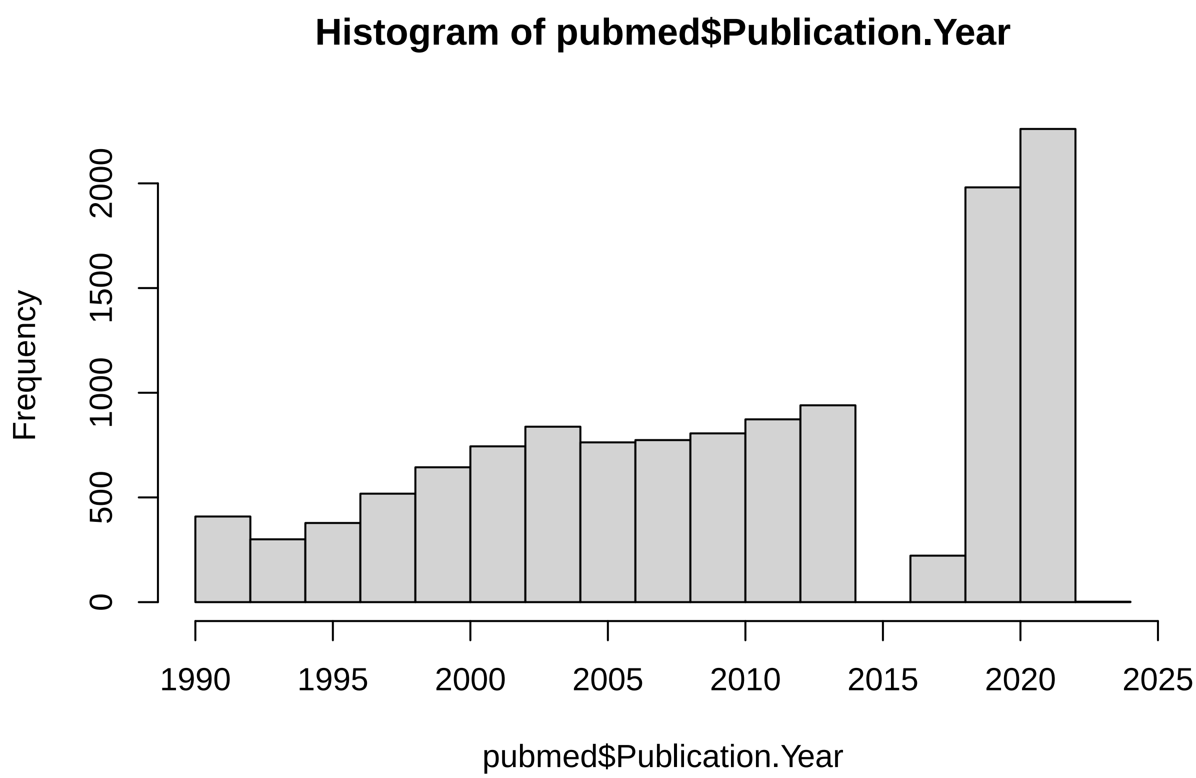
The data has been predownloaded as the Pubmed API via R is a bit slow.

Discussion points:

- Why has research publications increased dramatically in recent years? Is there something unsual with the dataset?
- What are the main cancer types being researched in Hong Kong?
- Is there any correlation between funding and cancer incidence and mortality?

Download HMRF grants and Pubmed data

```r
HMRF<- read.delim("https://github.com/StatBiomed/BMDS-book/raw/main/notebooks/module5-epidemi/HMRF_canc
#HMRF

pubmed<- read.delim("https://github.com/StatBiomed/BMDS-book/raw/main/notebooks/module5-epidemi/pubmed_
#pubmed

hist(pubmed$Publication.Year)
```

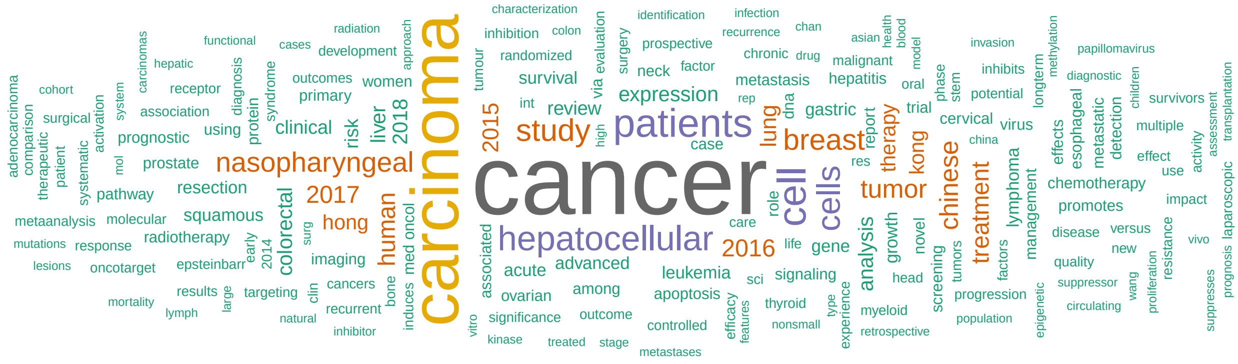


Make word cloud for grants

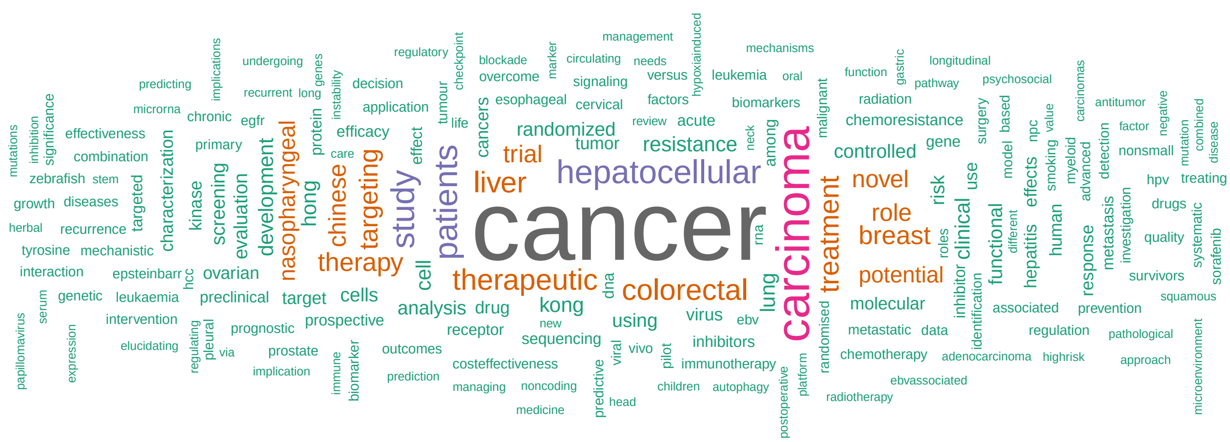
```
if (!require("ggwordcloud")) install.packages("ggwordcloud")
if (!require("tm")) install.packages("tm")
if (!require("dplyr")) install.packages("dplyr")
if (!require("RColorBrewer")) install.packages("RColorBrewer")
library(ggwordcloud)
library(tm)
library(dplyr)
library(RColorBrewer)

#Create a vector containing only the text
gtext <- HMRF$Project.Title
# Create a corpus
gdocs <- Corpus(VectorSource(gtext))
gdocs <- gdocs %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
gdocs <- tm_map(gdocs, content_transformer(tolower))
gdocs <- tm_map(gdocs, removeWords, stopwords("english"))

gdtm <- TermDocumentMatrix(gdocs)
gmatrix <- as.matrix(gdtm)
gwords <- sort(rowSums(gmatrix),decreasing=TRUE)
```

```r
gdf <- data.frame(word = names(gwords),freq=gwords)

set.seed(1235) # for reproducibility
gp<-ggwordcloud(words = gdf$word, freq = gdf$freq, min.freq = 1,max.words=200, random.order=FALSE, rot.p
#gp

#Create a vector containing only the text
pmtext <- pubmed$Title
# Create a corpus
pmdocs <- Corpus(VectorSource(pmtext))
pmdocs <- pmdocs %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
pmdocs <- tm_map(pmdocs, content_transformer(tolower))
pmdocs <- tm_map(pmdocs, removeWords, stopwords("english"))

pmdtm <- TermDocumentMatrix(pmdocs)
pmmatrix <- as.matrix(pmdtm)
pmwords <- sort(rowSums(pmmatrix),decreasing=TRUE)
pmdf <- data.frame(word = names(pmwords),freq=pmwords)

set.seed(1235) # for reproducibility
pm<-ggwordcloud(words = pmdf$word, freq = pmdf$freq, min.freq = 1,max.words=200, random.order=FALSE, ro
grid.arrange(arrangeGrob(pm, top = 'Pubmed'),arrangeGrob(gp, top = 'Grants'),nrow=2,heights=c(6,6))
```

Pubmed



Grants



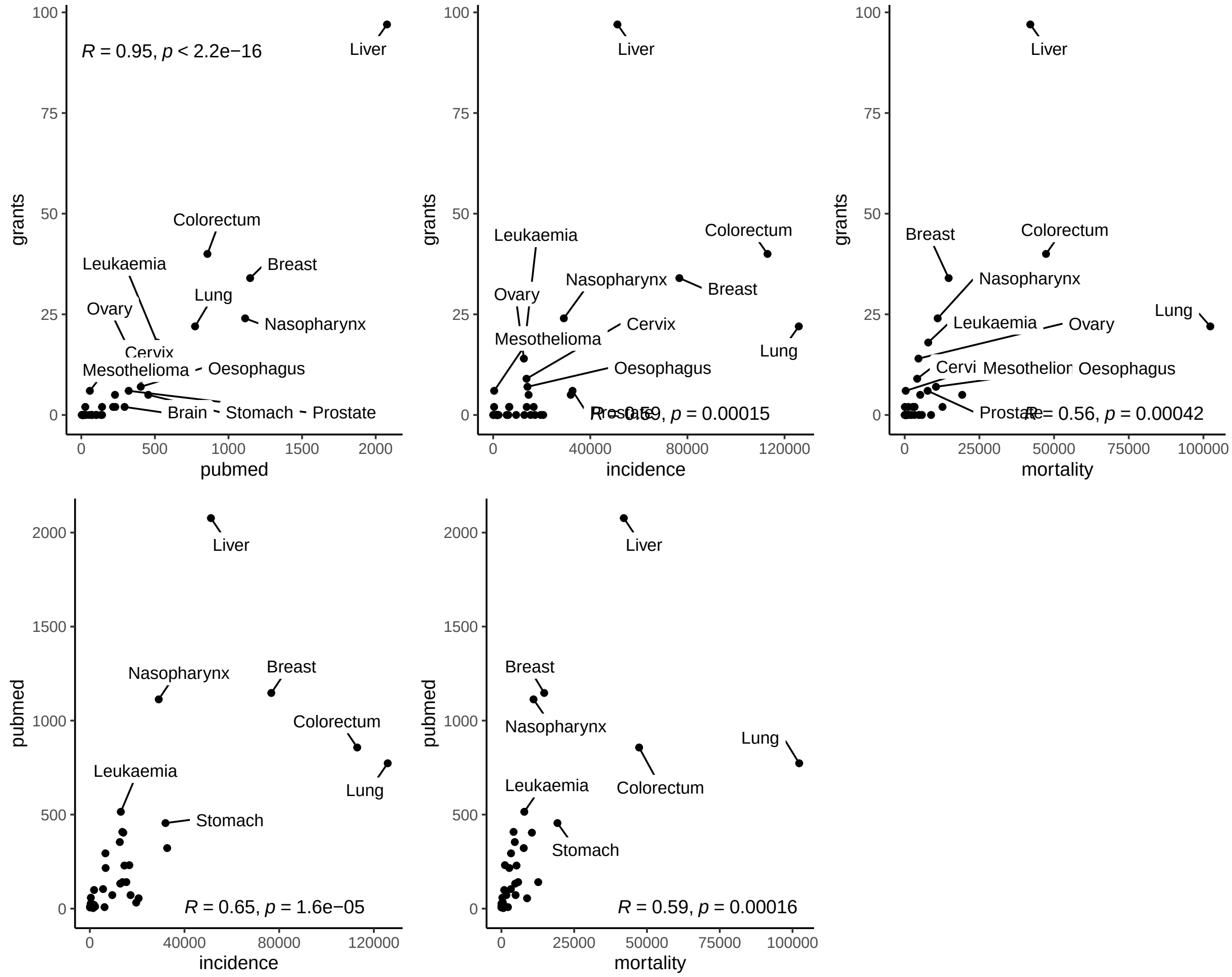
Make compare grant funding with incidence and mortality

```
if (!require("ggrepel")) install.packages("ggrepel")
if (!require("ggplot2")) install.packages("gglot2")
if (!require("gridExtra")) install.packages("gridExtra")
library(ggrepel)
library(ggplot2)
library(gridExtra)


grantpmsum<- read.delim("https://github.com/StatBiomed/BMDS-book/raw/main/notebooks/module5-epidemi/Gran
grantpmsum
```

```
##                  Cancer Grants Pubmed
## 1               Biliary      0     72
## 2               Bladder      0    141
## 3                 Brain      2    294
## 4                Breast     34   1147
## 5                Cervix      9    408
```

```
## 6            Colorectum     40    857
## 7                   Eye      2     27
## 8      Hodgkin lymphoma      0     21
## 9                Kaposi      0      6
## 10               Kidney      0    133
## 11               Larynx      0      8
## 12            Leukaemia     18    515
## 13                Liver     97   2077
## 14                 Lung     22    773
## 15             Melanoma      0     99
## 16         Mesothelioma      6     58
## 17     Multiple myeloma      0    104
## 18                Nasal      0     14
## 19          Nasopharynx     24   1113
## 20 Non-Hodgkin lymphoma      0     55
## 21                 Skin      0     32
## 22           Oesophagus      7    404
## 23                 Oral      5    229
## 24                Ovary     14    354
## 25             Pancreas      2    141
## 26                Penis      0      7
## 27             Placenta      0      9
## 28             Prostate      6    322
## 29              Sarcoma      2    216
## 30      Small intestine      0     11
## 31              Stomach      5    455
## 32               Testis      0     18
## 33               Thymus      0      3
## 34              Thyroid      2    231
## 35               Uterus      0     72
## 36                Vulva      0      8
```

```
HKCancer_inc <- HKCancer[HKCancer$Type=='incidence',]
HKCancer_inc_sum <- data.frame(incidence=colSums(HKCancer_inc[,-(1:4)]))

HKCancer_mort <- HKCancer[HKCancer$Type=='mortality',]
HKCancer_mort_sum <- data.frame(mortality=colSums(HKCancer_inc[,-(1:4)]))

HKCancer_compare <-data.frame(cancer=grantpmsum$Cancer, grants=grantpmsum$Grants, pubmed=grantpmsum$Pubm

gp<-ggplot(HKCancer_compare, aes(y=grants, x=pubmed)) + geom_point() + scale_color_brewer(palette="Dark2

pinc<-ggplot(HKCancer_compare, aes(y=pubmed, x=incidence)) + geom_point() + scale_color_brewer(palette=

pmor<-ggplot(HKCancer_compare, aes(y=pubmed, x=mortality)) + geom_point() + scale_color_brewer(palette=


ginc<-ggplot(HKCancer_compare, aes(y=grants, x=incidence)) + geom_point() + scale_color_brewer(palette=

gmor<-ggplot(HKCancer_compare, aes(y=grants, x=mortality)) + geom_point() + scale_color_brewer(palette=

grid.arrange(gp,ginc,gmor,pinc, pmor, ncol=3)
```



5. Task (to be completed in Tutotial 2 by 3:00pm and shared with other tutorial group)

As a group, discuss what cancer type you think is most worthy of funding in Hong Kong. Prepare 2-3 PowerPoint/Google Slides summarising your decision. Include some points and figures supporting your final conclusion. If possible also include suggestions of what other data/analyses can be performed and/or other cancer types that are also in need of funding in Hong Kong.