

SEERaBomb Overview

Tom Radivoyevitch

April 16, 2014

Introduction

SEERaBomb is for SEER and Japanese A-bomb survivor data analysts. It contributes speed to SEER analyses by reducing file sizes to contain only items of interest. To obtain the data please visit the links in [gettingData.pdf](#). Use cases are given in R scripts in the courses and papers directories.

SEER Data R Binaries

The incidence directory of the SEER dataset contains a SAS file that defines the field names, their starting positions, and their fixed widths. This file can be used to read the SEER data into SAS, but it is used here to: 1) present the field choices (see [fieldNames.html](#) and the output of `getFields()`); and 2) given user choices, automatically determine the sequence of widths needed to extract the data of interest using the speedy R package LaF. `getFields()` has one parameter, `seerHome="~/data/SEER"`, which should be overridden if the SEER data lives elsewhere. Its data.frame output and the SEER file `seerdic.pdf` in the SEER incidence directory must be thoroughly examined to determine which fields will be useful. Once this is determined, the output and list of field choices, the default of which is

```
picks=c("casenum","reg","race","sex","agedx","yrbrth",  
        "seqnum","yrdx","histo2","histo3","radiatn","agerec",  
        "ICD9","histrec","numprims","COD","surv"),
```

must then be inputted into `pickFields()`.

The output of `pickFields()` contains not only pulled rows from the input, but also inserted rows with widths computed to fill the gaps of no interest. Knowing these gap sizes enables fast file reading by LaF in `mkSEER()`. This function produces R Data binaries in SEER dataset subdirectories of `seerHome` such as `~/data/SEER/00` for SEER18 data (which was collected since 2000).

This work was supported by the National Cancer Institute and Tufts Integrative Cancer Biology Program under U54CA149233-029689.

```

library(SEERaBomb)
df = getFields()
(df = pickFields(df))

##          start width  names          desc      type
## casenum      1     8 casenum      Patient ID number integer
## reg          9    10 reg          Registry ID integer
## 3           19     1          string
## race         20     2 race          Race/Ethnicity integer
## 5           22     2          string
## sex          24     1 sex          Sex integer
## agedx        25     3 agedx        Age at diagnosis integer
## yrbrth       28     4 yrbrth       Year of birth integer
## 9           32     3          string
## seqnum       35     2 seqnum       Sequence Number--Central integer
## 11          37     2          string
## yrdx         39     4 yrdx         Year of diagnosis integer
## 13          43     5          string
## histo2       48     4 histo2       Histology (92-00) ICD-0-2 integer
## 15          52     1          string
## histo3       53     4 histo3       Histologic Type ICD-0-3 integer
## 17          57    110          string
## radiatn     167     1 radiatn      RX Summ--Radiation integer
## 19          168    24          string
## agerec       192     2 agerec       Age Recode <1 Year olds integer
## 21          194    10          string
## ICD9         204     4 ICD9        Recode ICD-0-2 to 9 integer
## 23          208    18          string
## histrec      226     2 histrec     Histology Recode--Broad Groupings integer
## 25          228    15          string
## numprims     243     2 numprims    Number of primaries integer
## 27          245    10          string
## COD          255     5 COD Cause of death to SEER site recode integer
## 29          260    41          string
## surv         301     4 surv        Survival months integer
## 31          305    27          string

```